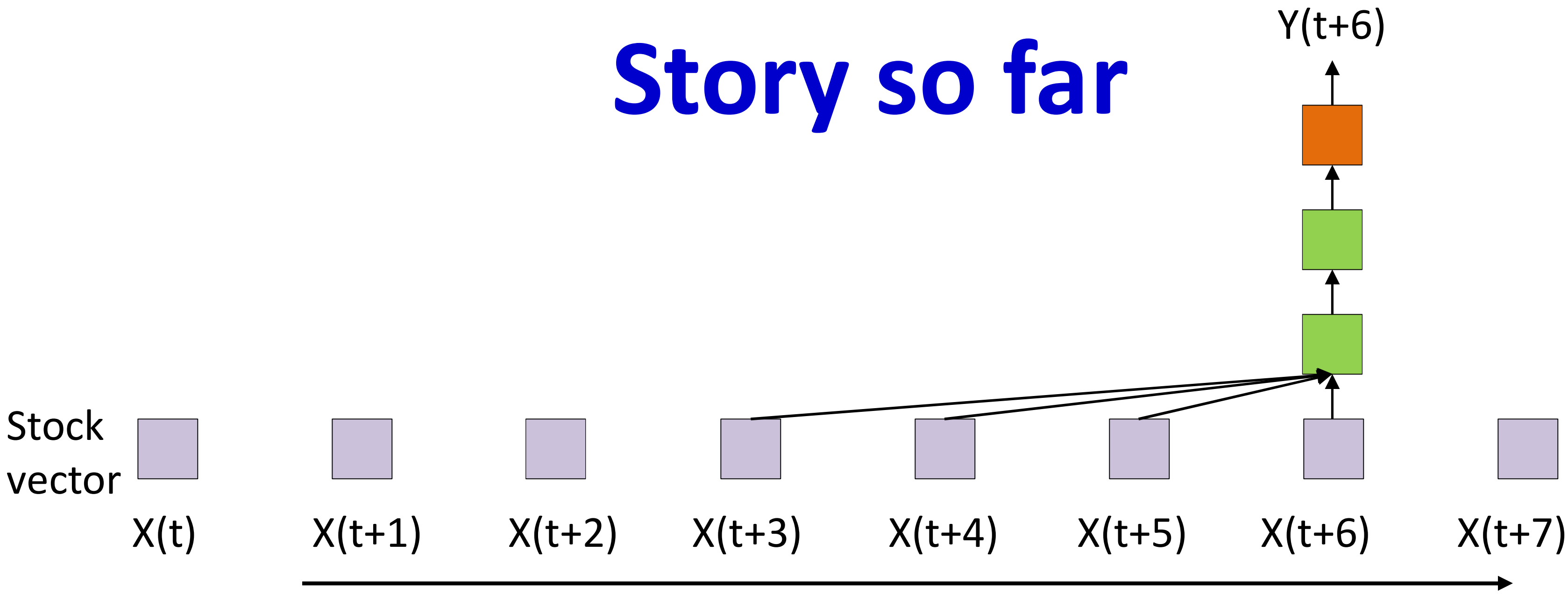
Deep Learning
Recurrent Networks
Part 3

Recap: Recurrent networks can be incredibly effective

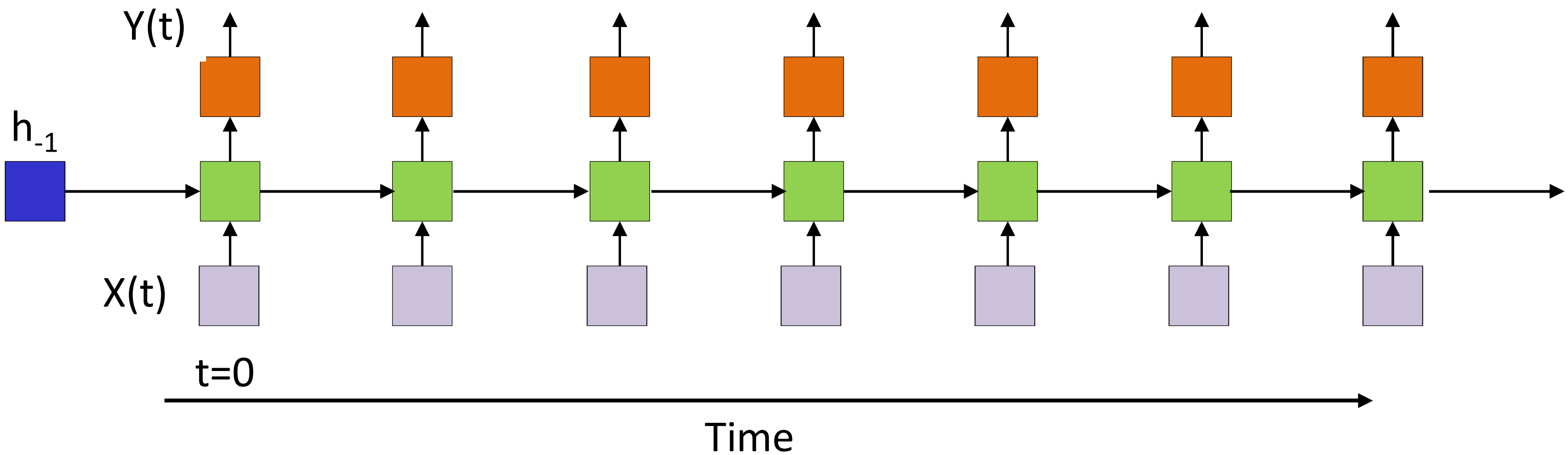
```
/*
 * Increment the size file of the new incorrect UI_FILTER group information
 * of the size generatively.
 */
static int indicate_policy(void)
{
  int error;
  if (fd == MARN_EPT) {
    /*
     * The kernel blank will coeld it to userspace.
     */
    if (ss->segment < mem_total)
      unblock_graph_and_set_blocked();
    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_SECON
  return segtable;
}
```

Story so far



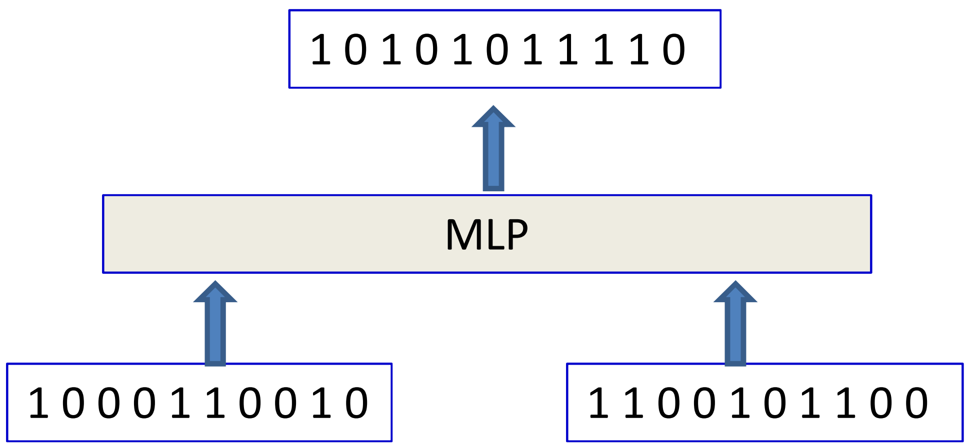
- ***Iterated structures*** are good for analyzing time series data with short-time dependence on the past
  - These are "***Time delay***" neural nets, AKA ***convnets***

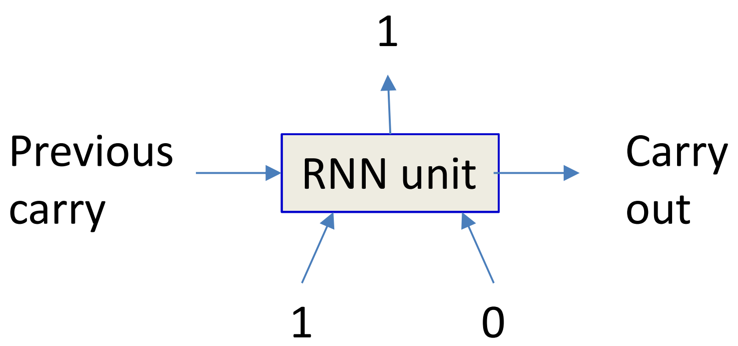
Story so far



- Iterated structures are good for analyzing time series data with short-time dependence on the past
  - These are “Time delay” neural nets, AKA convnets
- ***Recurrent structures*** are good for analyzing time series data with ***long-term*** dependence on the past
  - These are ***recurrent*** neural networks

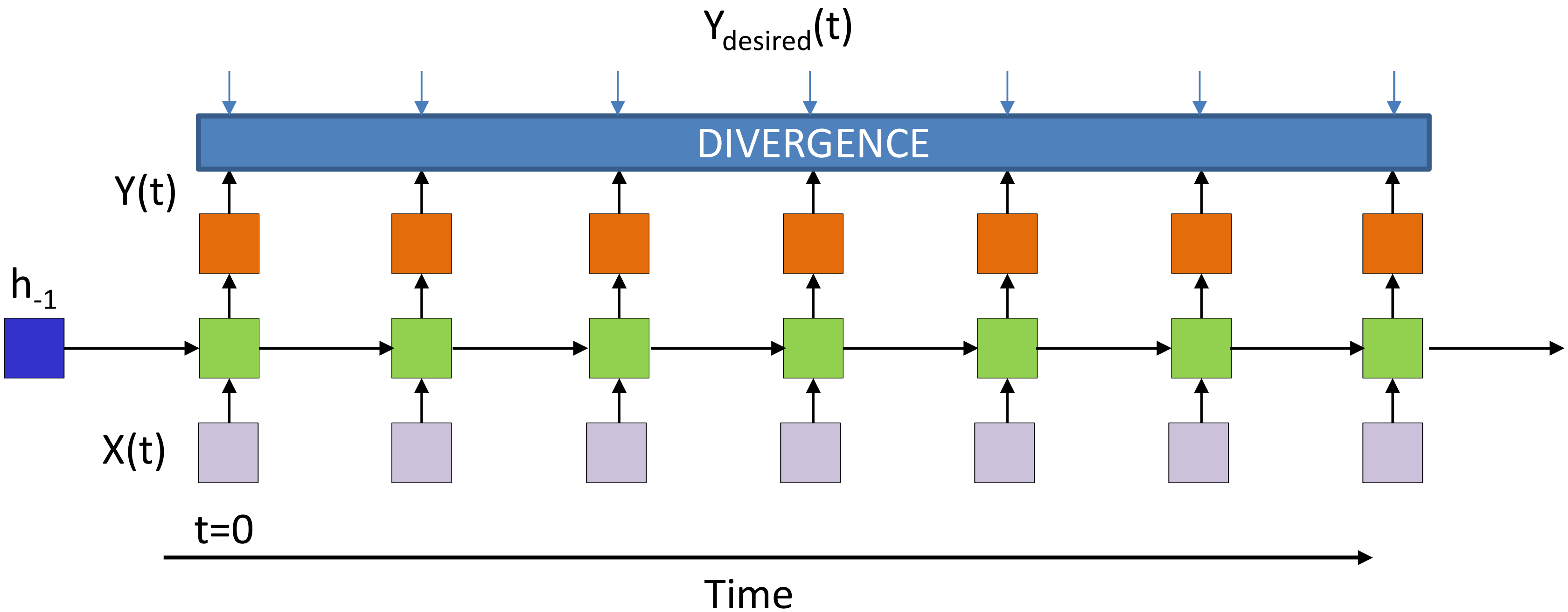
Recurrent structures can do what static structures cannot





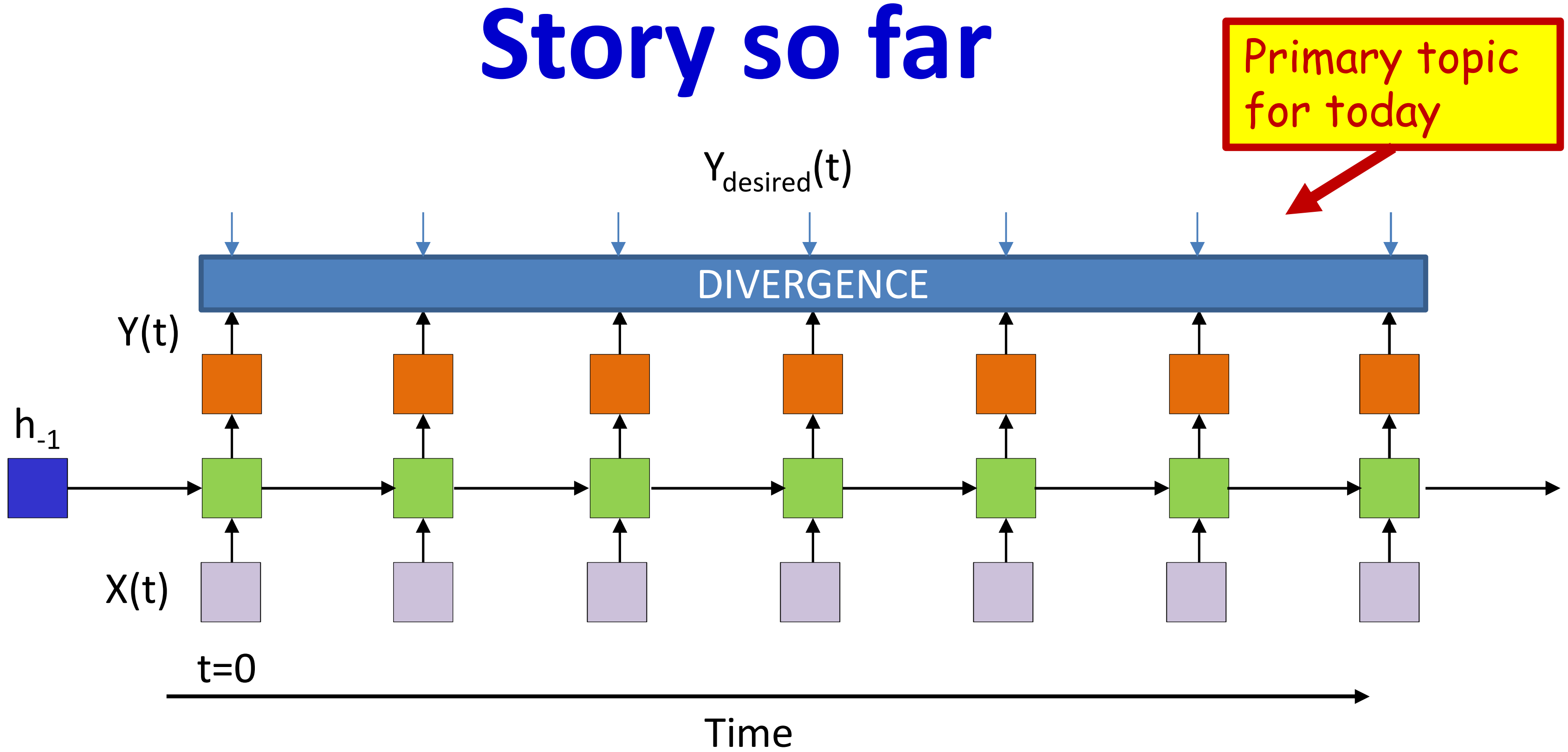
- The addition problem: Add two N-bit numbers to produce a N+1-bit number
  - Input is binary
  - Will require large number of training instances
    - Output must be specified for every pair of inputs
    - Weights that generalize will make errors
  - Network trained for N-bit numbers will not work for N+1 bit numbers
- An RNN learns to do this very quickly
  - With very little training data!

Story so far



- Recurrent structures can be trained by minimizing the divergence between the ***sequence*** of outputs and the ***sequence*** of desired outputs
  - Through gradient descent and backpropagation

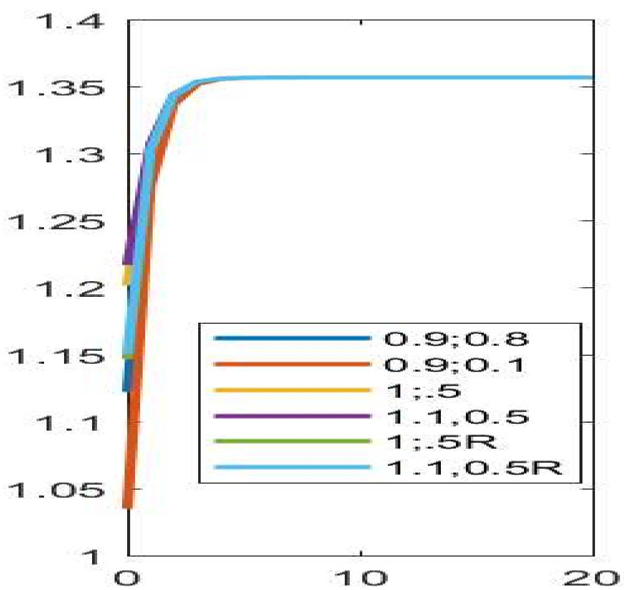
Story so far



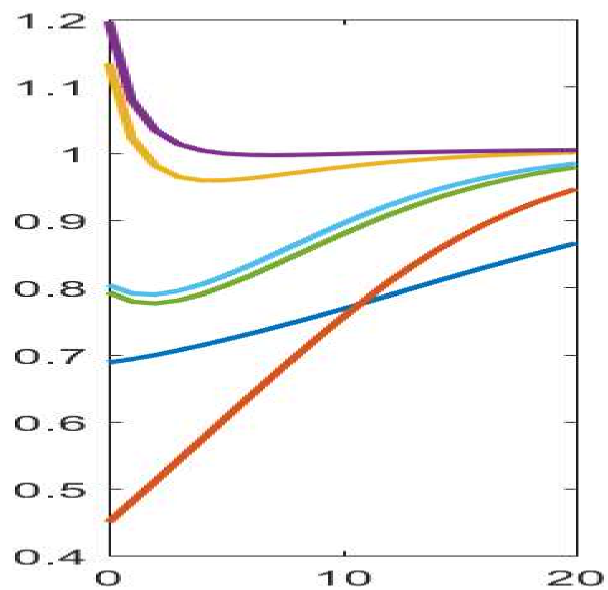
- Recurrent structures can be trained by minimizing the divergence between the ***sequence*** of outputs and the ***sequence*** of desired outputs
  - Through gradient descent and backpropagation

Story so far: stability

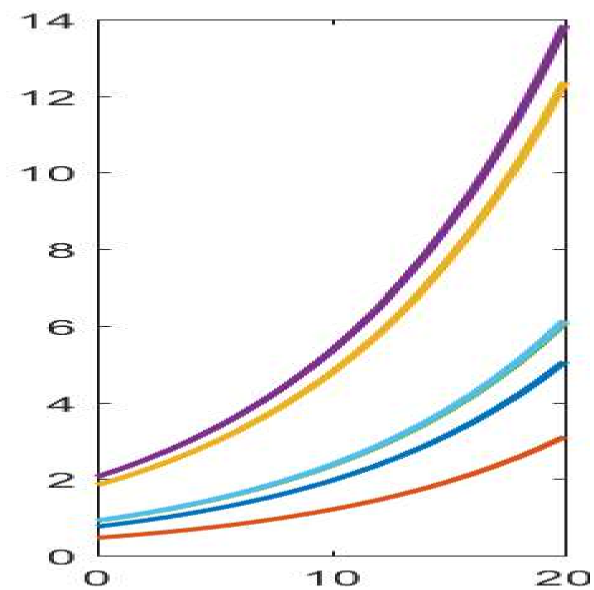
- Recurrent networks can be unstable
  - And not very good at remembering at other times



sigmoid



tanh

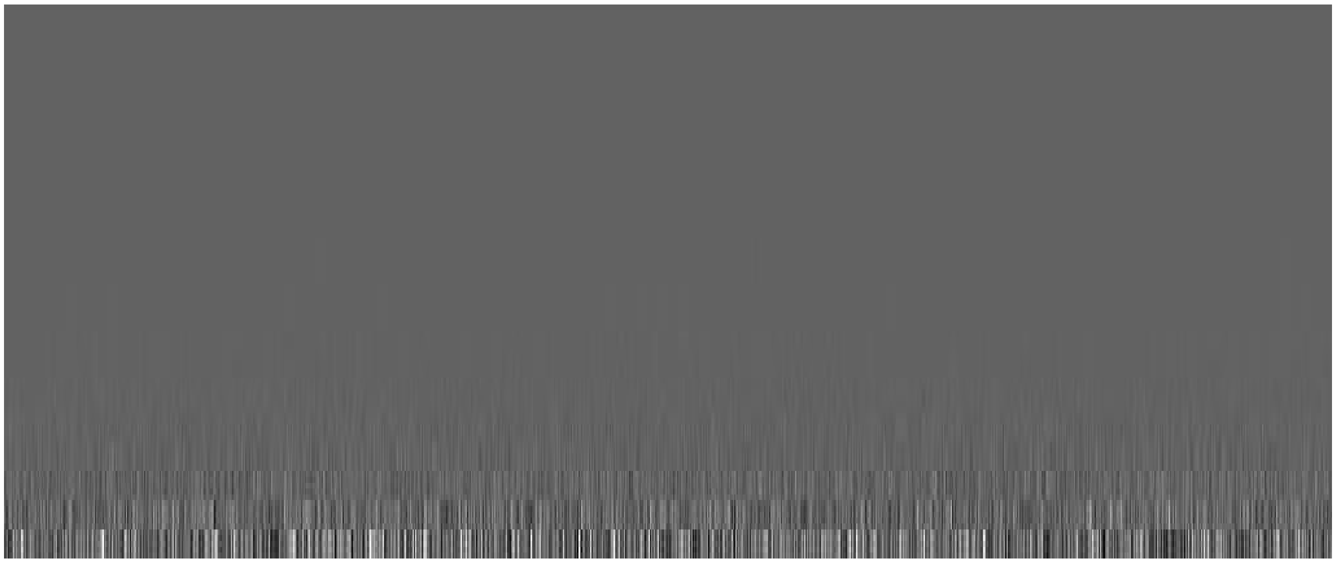


relu

Vanishing gradient examples..

ELU activation, Batch gradients

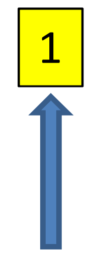
Input layer



Output layer

- Learning is difficult: gradients tend to vanish..

The long-term dependency problem

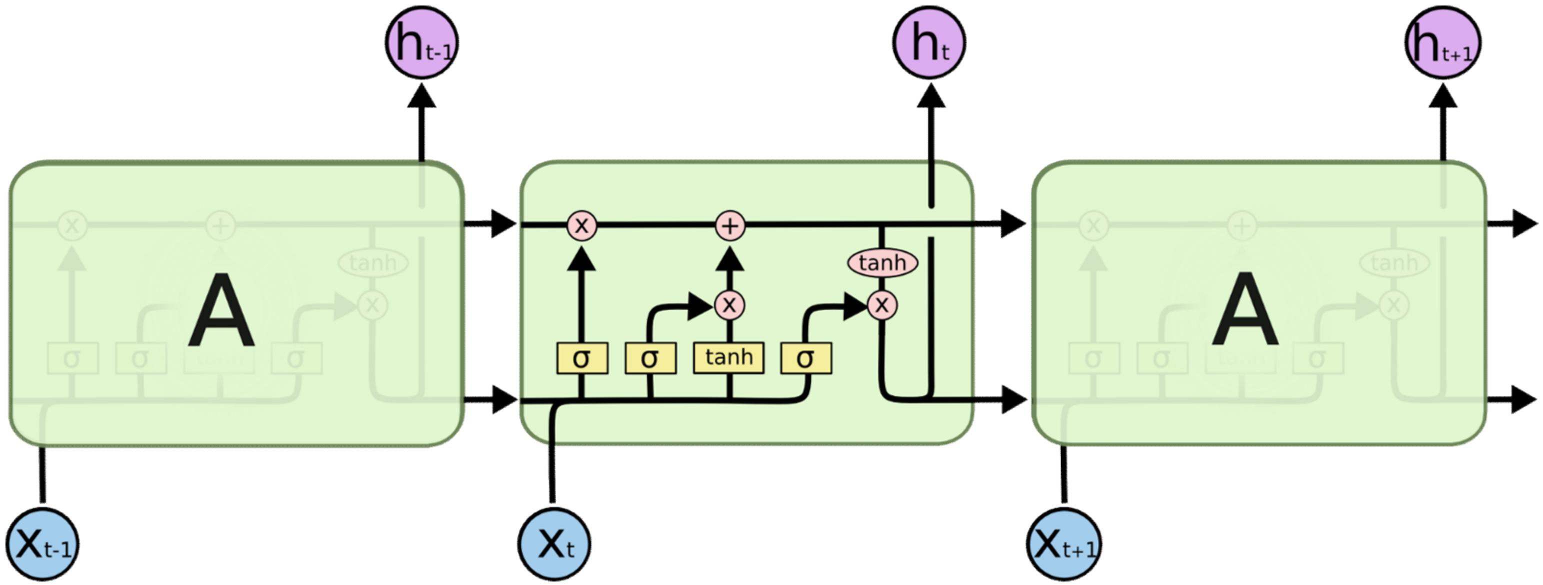


PATTERN1 [..............................] PATTERN 2

Jane had a quick lunch in the bistro. Then *she..*

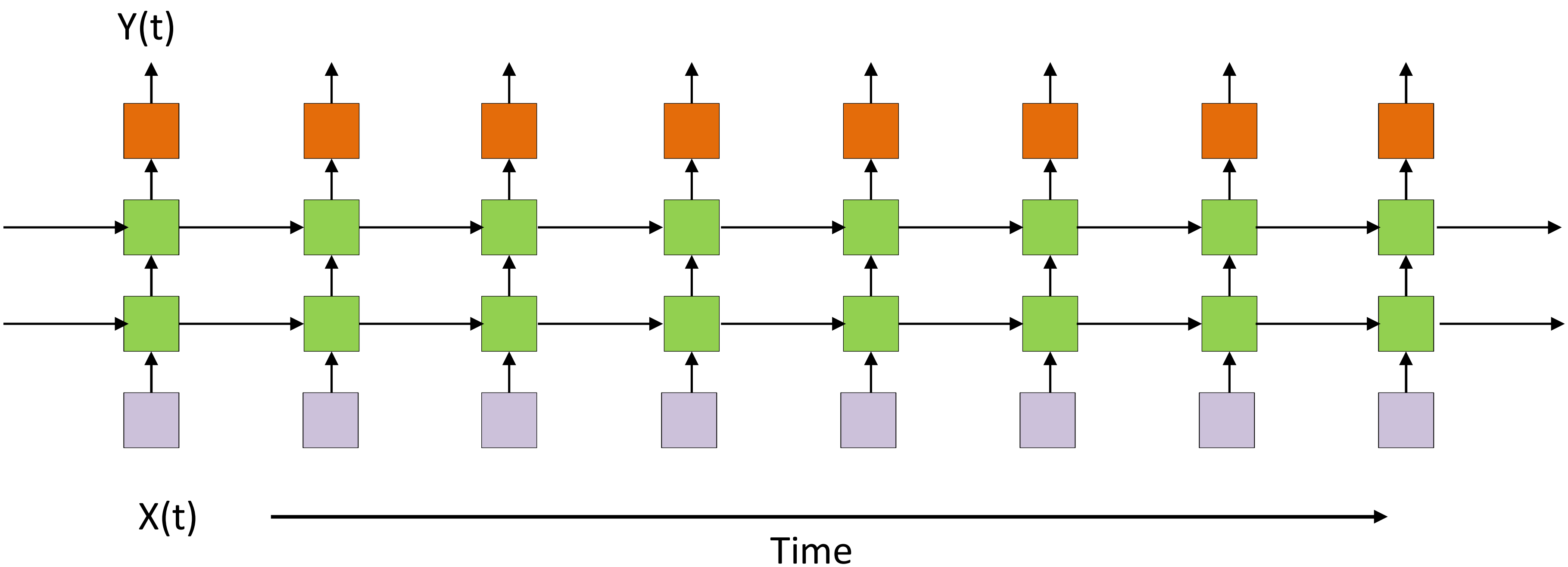
- Long-term dependencies are hard to learn in a network where memory behavior is an untriggered function of the *network*
  - Need it to be a triggered response to *input*

Long Short-Term Memory



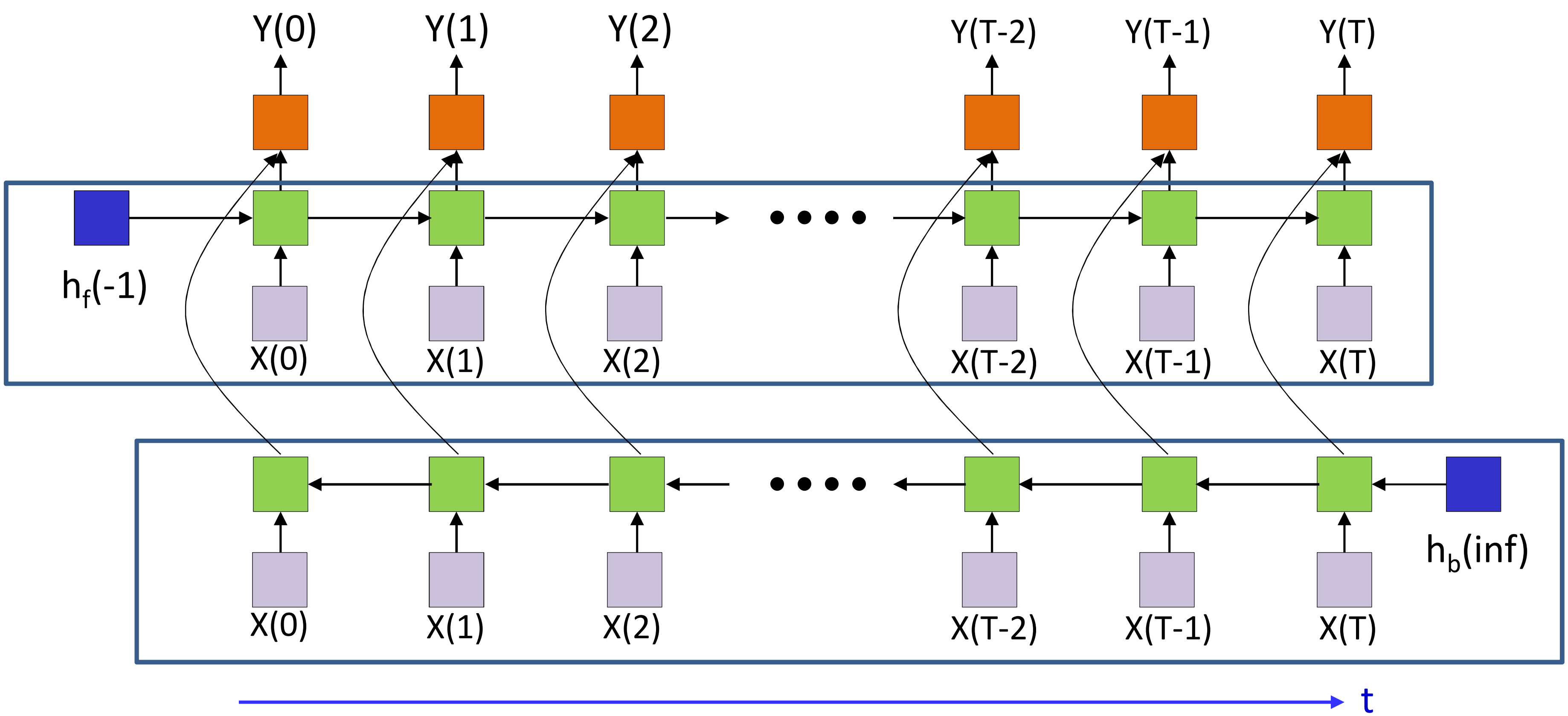
- The LSTM addresses the problem of *input-dependent* memory behavior

LSTM-based architecture



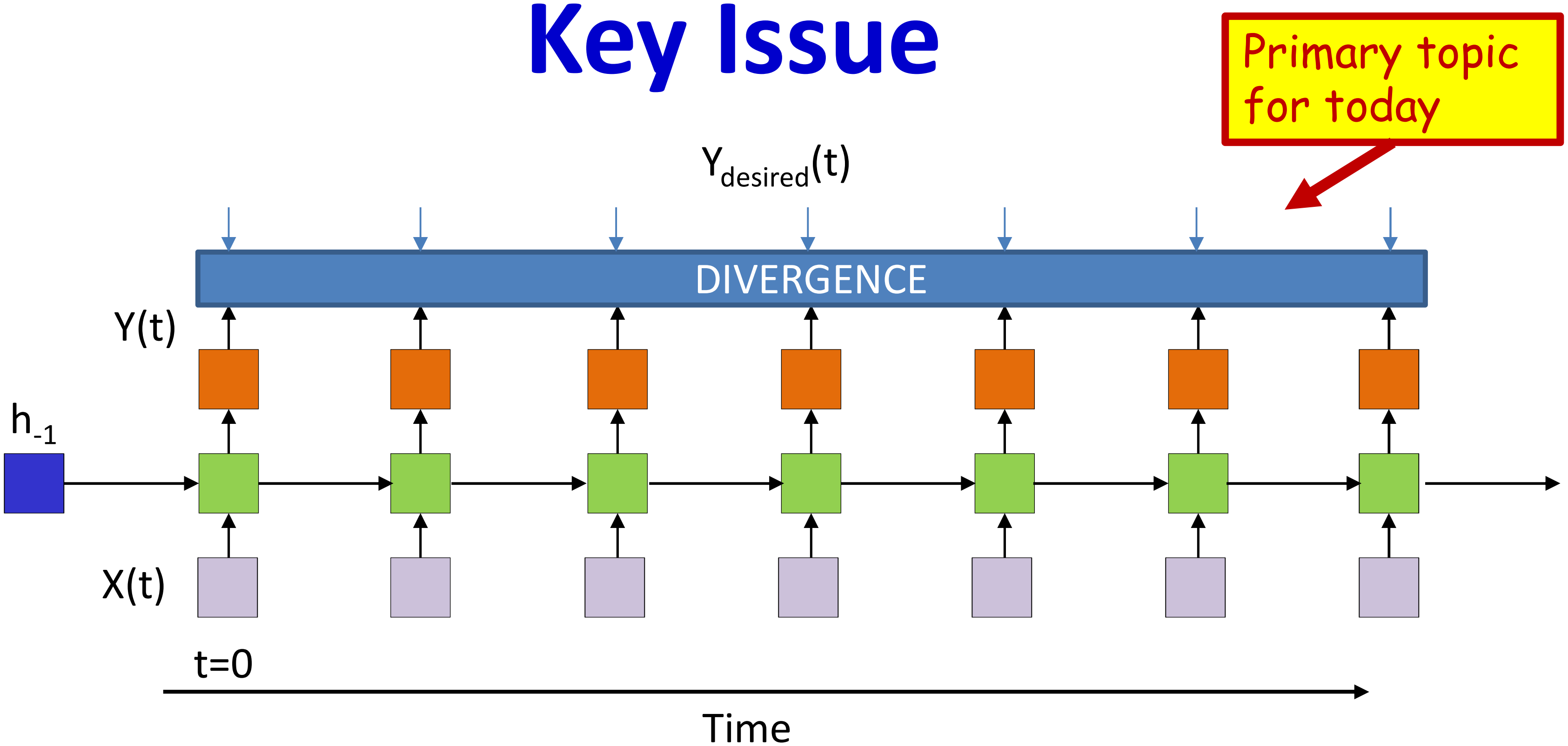
- LSTM based architectures are identical to RNN-based architectures

Bidirectional LSTM



- Bidirectional version..

Key Issue



- How do we define the divergence
- Also: how do we compute the outputs..

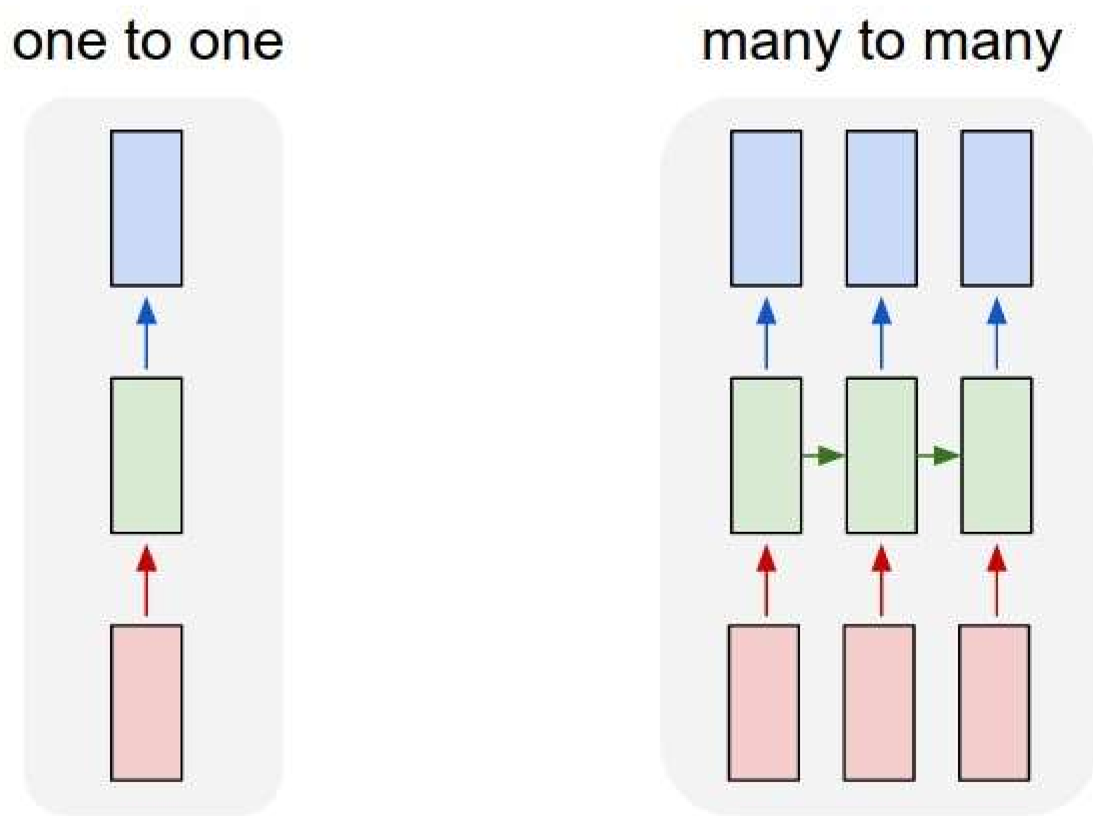
What follows in this series on recurrent nets

- Architectures: How to train recurrent networks of different architectures

- Synchrony: How to train recurrent networks when
  - The target output is time-synchronous with the input
  - The target output is order-synchronous, but not time synchronous
  - Applies to only some types of nets

- How to make predictions/inference with such networks

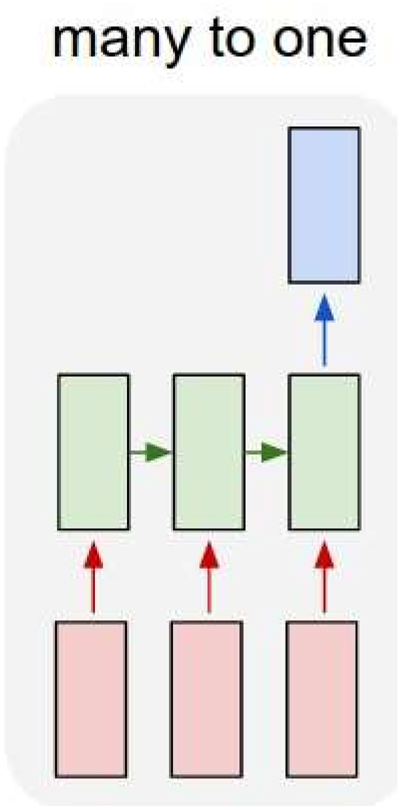
Variants on recurrent nets

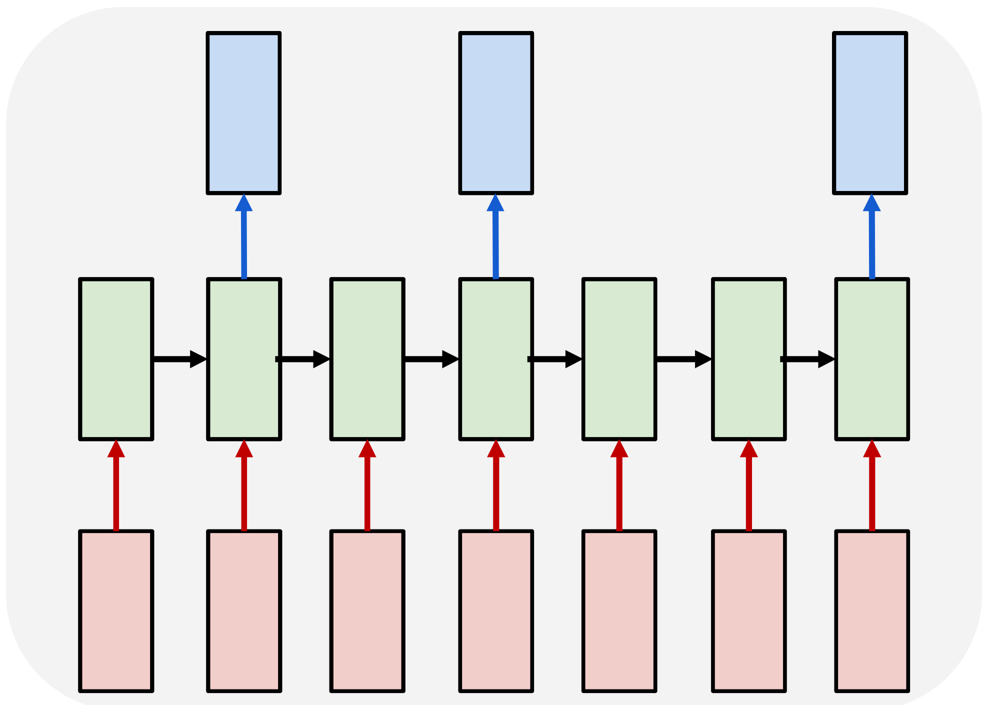


Images from Karpathy

- Conventional MLP
- Time-synchronous outputs
  - E.g. part of speech tagging

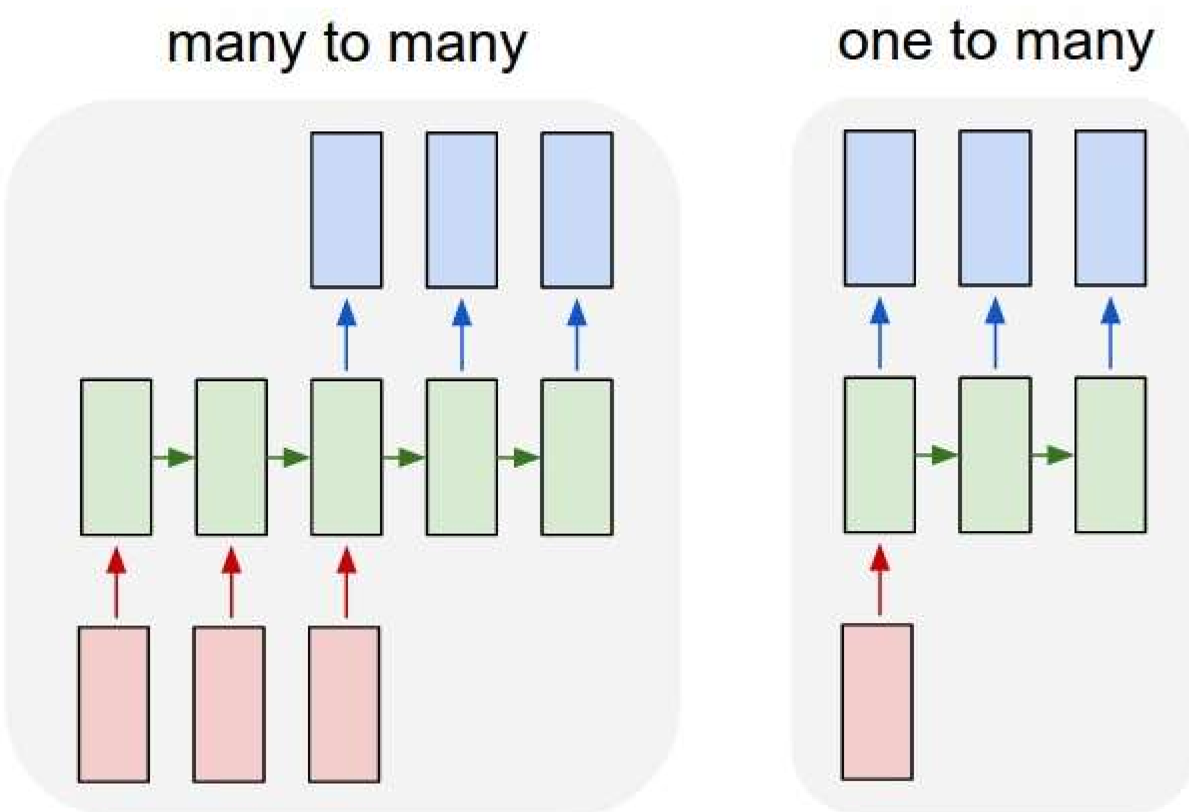
Variants on recurrent nets





- Sequence classification: Classifying a full input sequence
  - E.g phoneme recognition
- Order synchronous , time asynchronous sequence-to-sequence generation
  - E.g. speech recognition
  - Exact location of output is unknown a priori

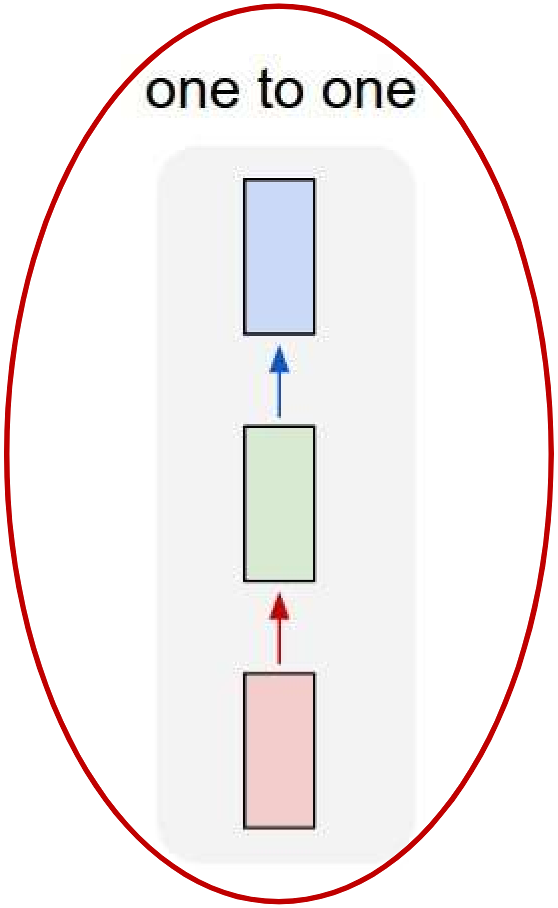
Variants

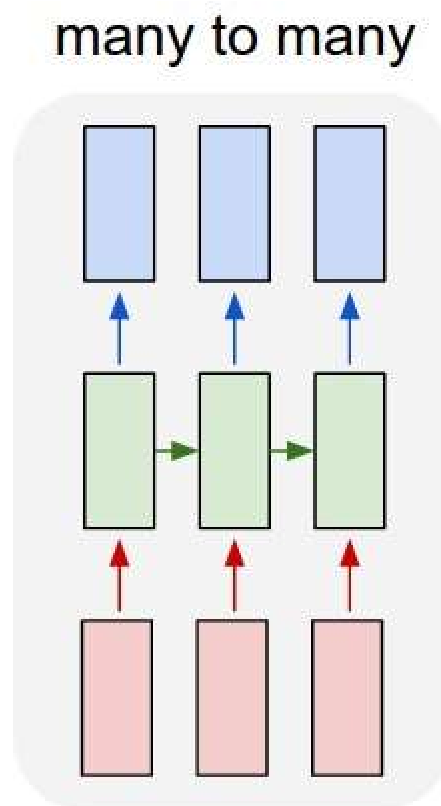


Images from Karpathy

- A posteriori sequence to sequence: Generate output sequence after processing input
  - E.g. language translation
- Single-input a posteriori sequence generation
  - E.g. captioning an image

Variants on recurrent nets

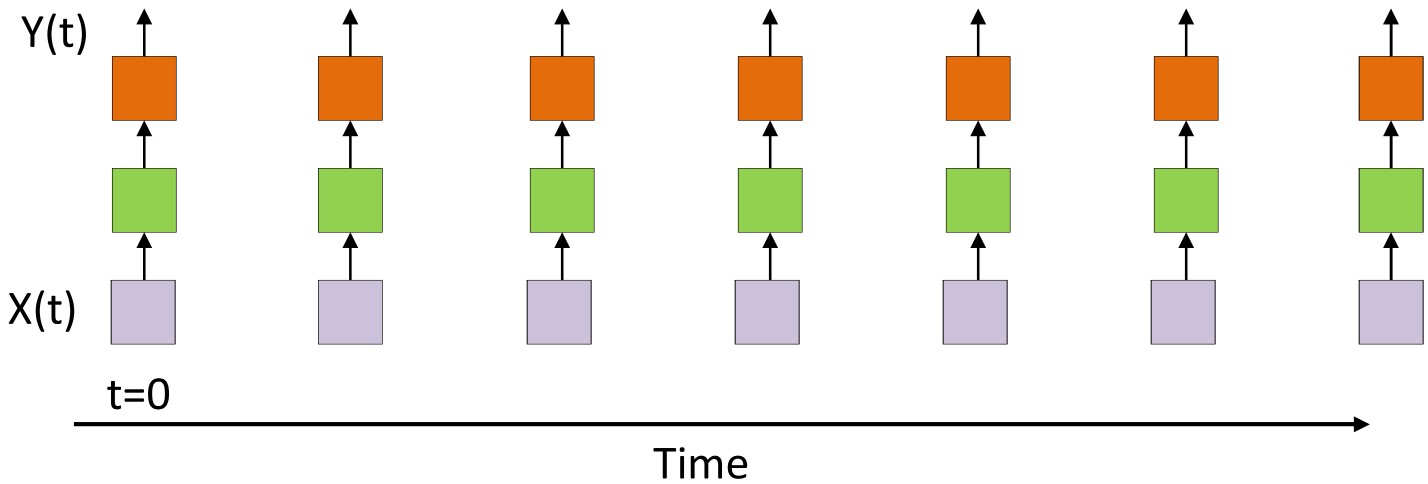




Images from Karpathy

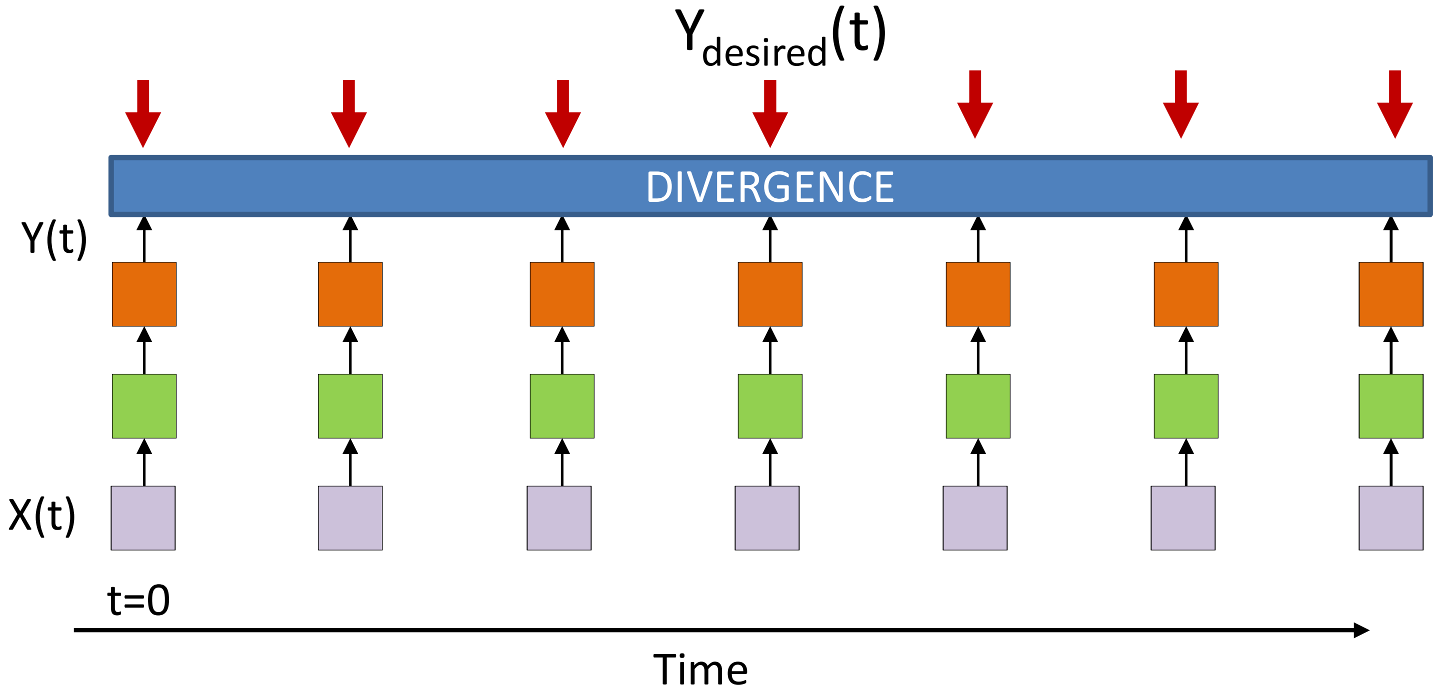
- Conventional MLP
- Time-synchronous outputs
  - E.g. part of speech tagging

Regular MLP for processing sequences



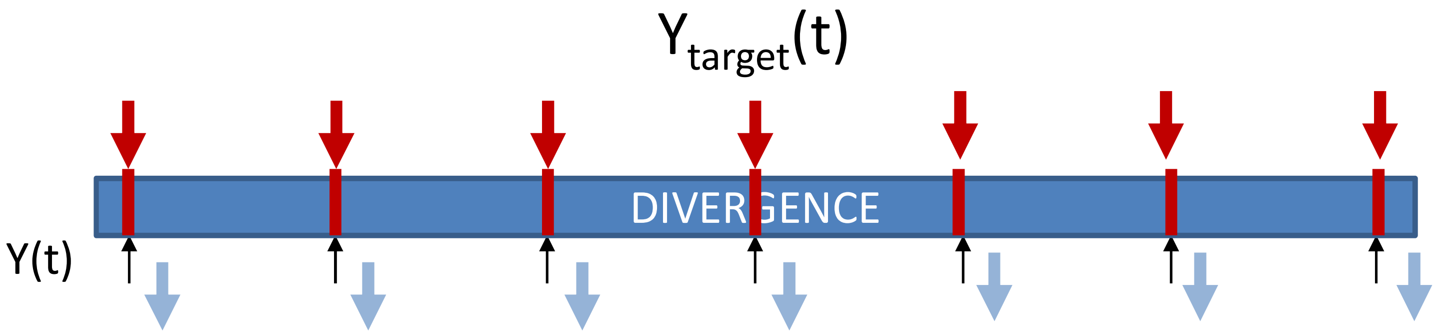
- No recurrence in model
  - Exactly as many outputs as inputs
  - Every input produces a unique output

Learning in a Regular MLP



- No recurrence
  - Exactly as many outputs as inputs
    - **One to one correspondence between desired output and actual output**
  - The output at time t is not a function of the output at $t' \neq t$.

Regular MLP



- Gradient backpropagated at each time

$$\nabla_{Y(t)} Div\left(Y_{target}(1 \ldots T), Y(1 \ldots T)\right)$$

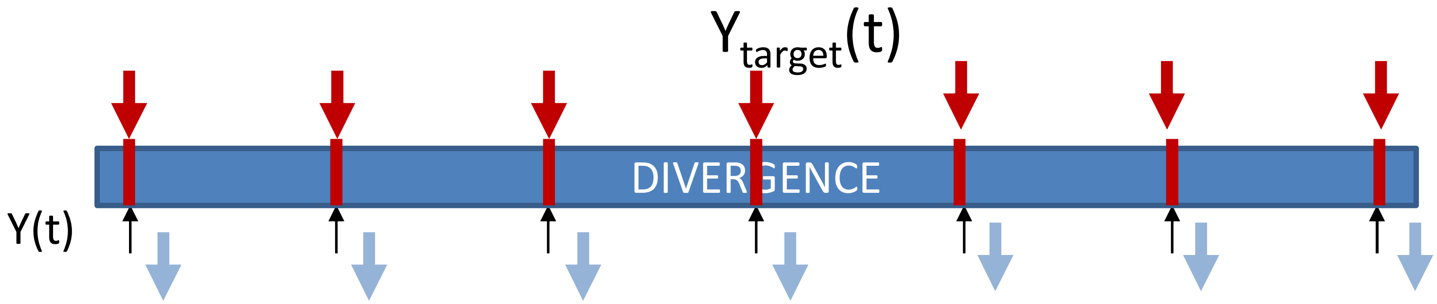
- Common assumption:

$$Div\left(Y_{target}(1 \ldots T), Y(1 \ldots T)\right) = \sum_t w_t Div\left(Y_{target}(t), Y(t)\right)$$

$$\nabla_{Y(t)} Div\left(Y_{target}(1 \ldots T), Y(1 \ldots T)\right) = w_t \nabla_{Y(t)} Div\left(Y_{target}(t), Y(t)\right)$$

  - w_t is typically set to 1.0
  - This is further backpropagated to update weights etc

Regular MLP



- Gradient backpropagated at each time

$$\nabla_{Y(t)} Div\big(Y_{target}(1 \dots T), Y(1 \dots T)\big)$$

- Common assumption:

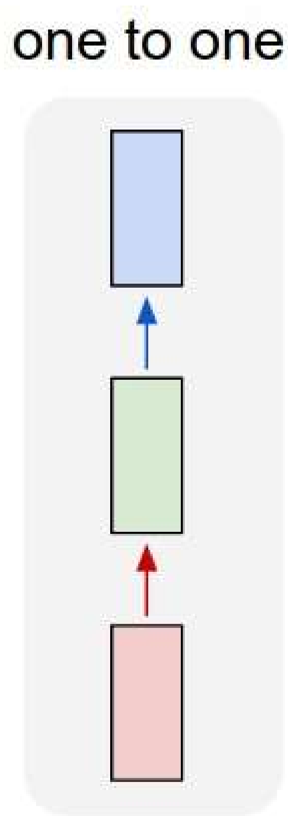
$$Div\big(Y_{target}(1 \dots T), Y(1 \dots T)\big) = \sum_t Div\big(Y_{target}(t), Y(t)\big)$$

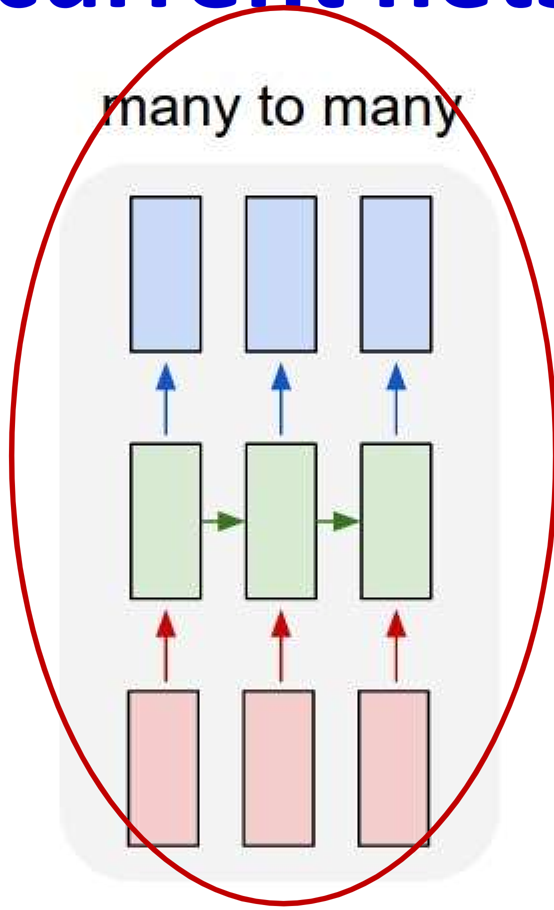
$$\nabla_{Y(t)} Div\big(Y_{target}(1 \dots T), Y(1 \dots T)\big) = \nabla_{Y(t)} Div\big(Y_{target}(t), Y(t)\big)$$

  – This is further backpropagated to update weights etc

Typical Divergence for classification: $Div\big(Y_{target}(t), Y(t)\big) = Xent(Y_{target}, Y)$

Variants on recurrent nets

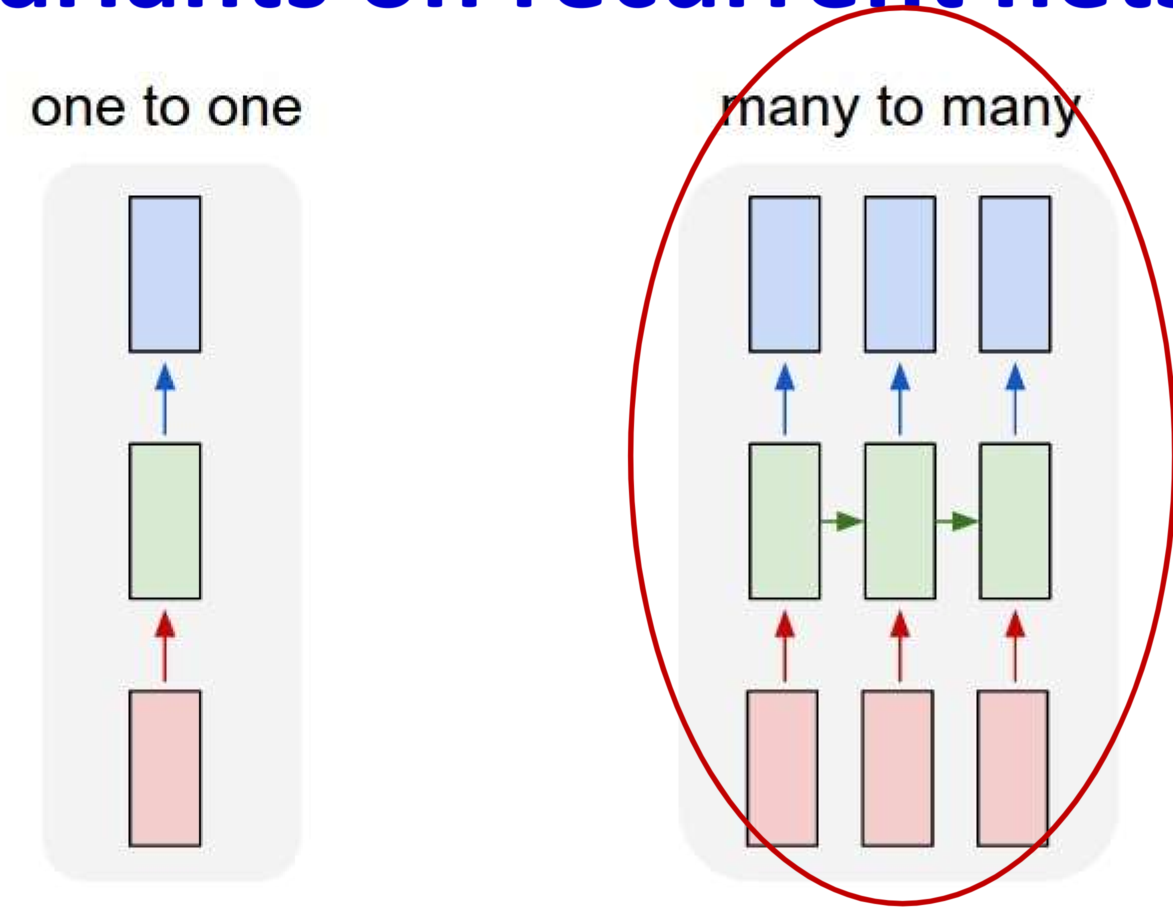




Images from Karpathy

- Conventional MLP
- Time-synchronous outputs
  - E.g. part of speech tagging

Variants on recurrent nets

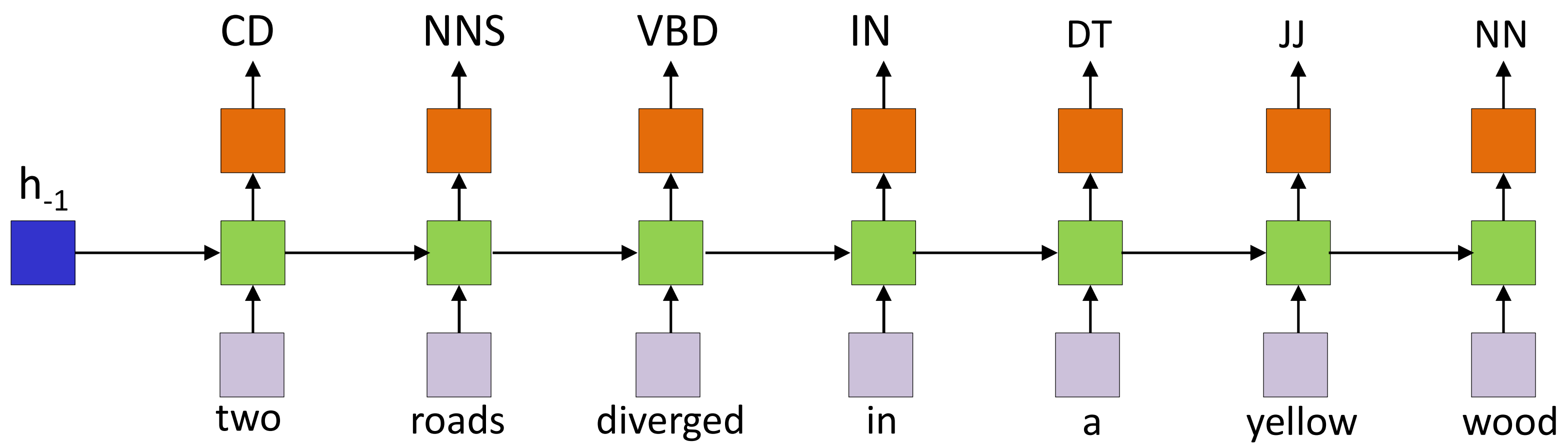


Images from Karpathy

With a brief detour into modelling language

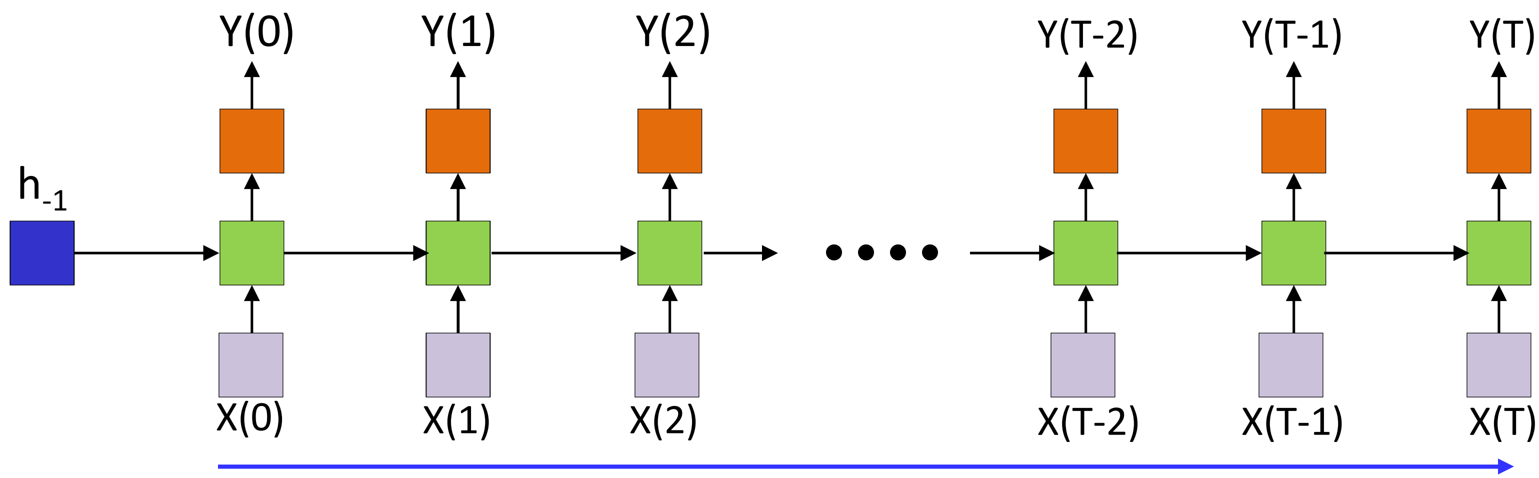
- Time-synchronous outputs
  - E.g. part of speech tagging

Time synchronous network



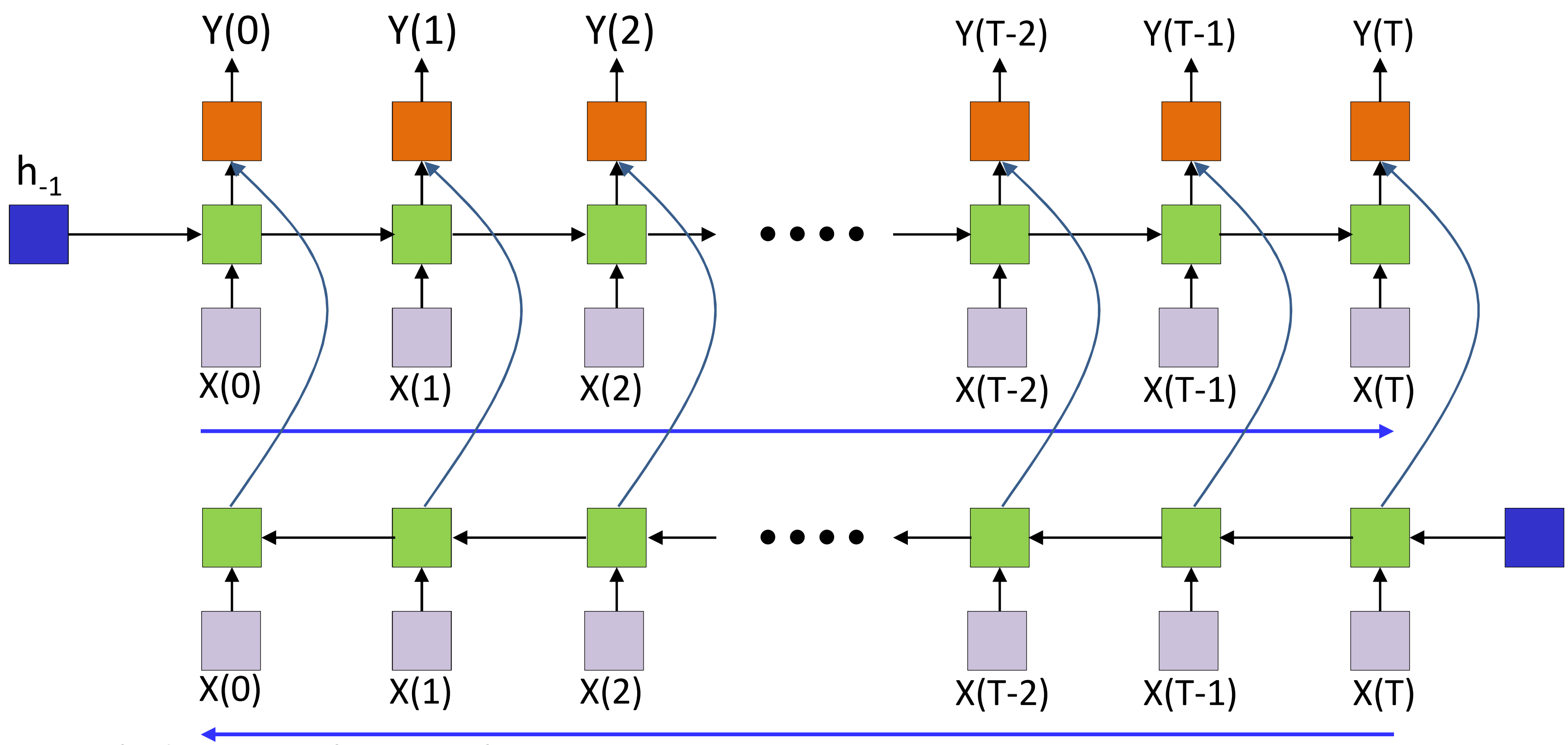
- Network produce one output for each input
  - With one-to-one correspondence
  - E.g. Assigning grammar tags to words
    - May require a bidirectional network to consider both past and future words in the sentence

Time-synchronous networks: Inference



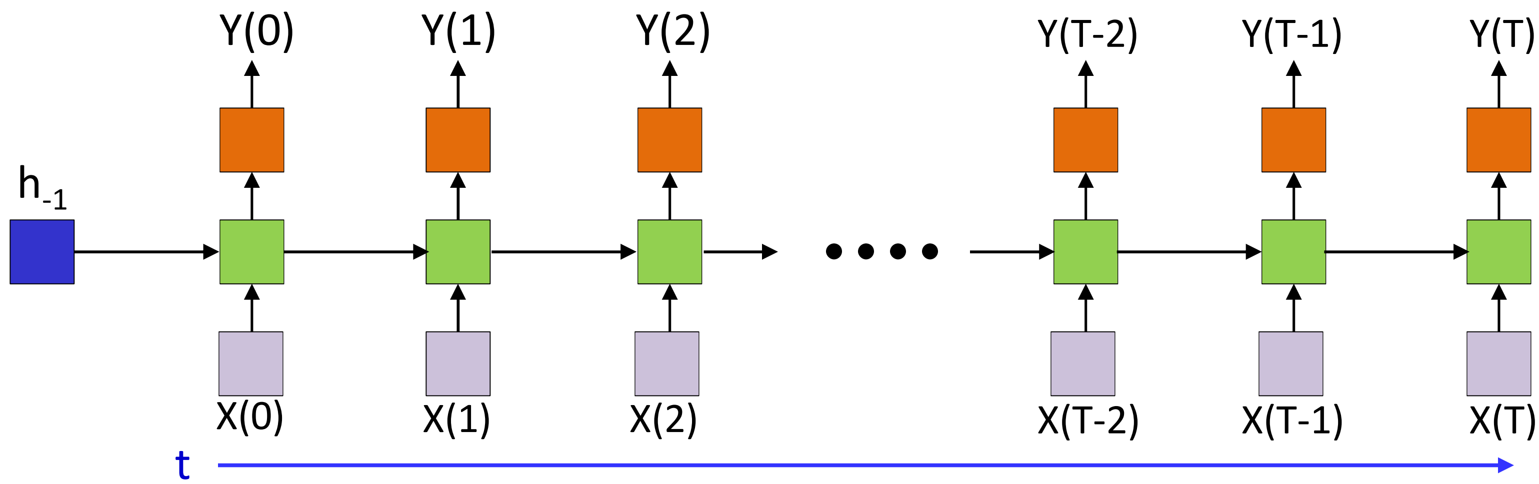
- Process input left to right and produce output after each input

Time-synchronous networks: Inference



- For bidirectional networks:
  - Process input left to right using forward net
  - Process it right to left using backward net
  - Combine their hidden outputs to produce one output per input symbol
- Rest of the lecture(s) will not specifically consider bidirectional nets, but the discussion generalizes

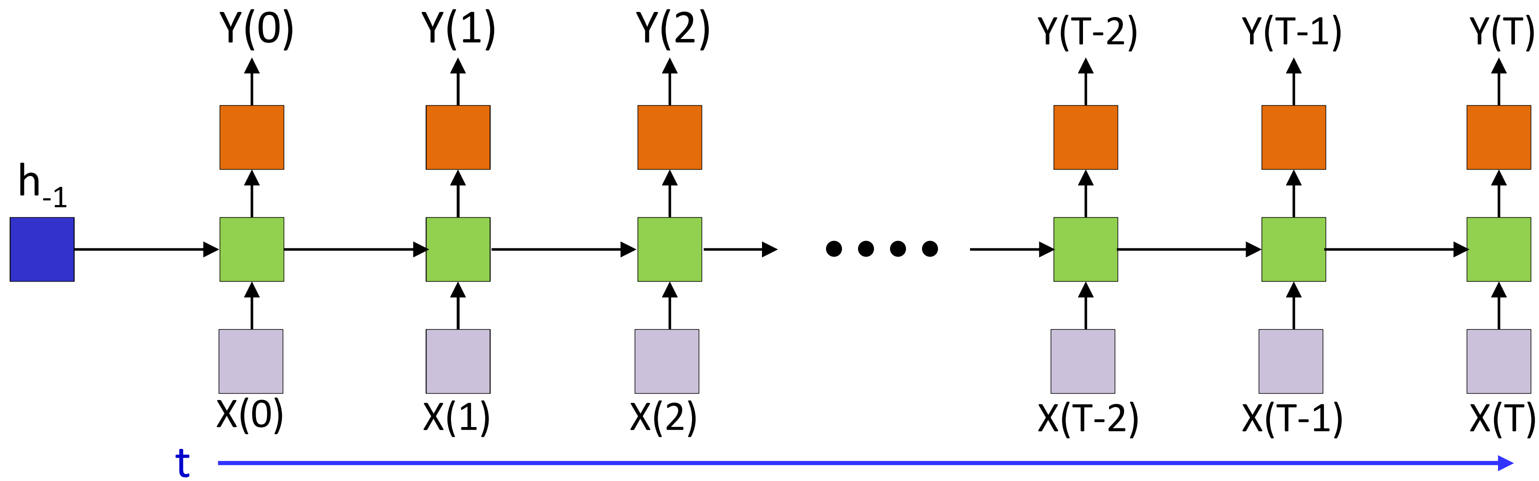
How do we *train* the network



- Back propagation through time (BPTT)

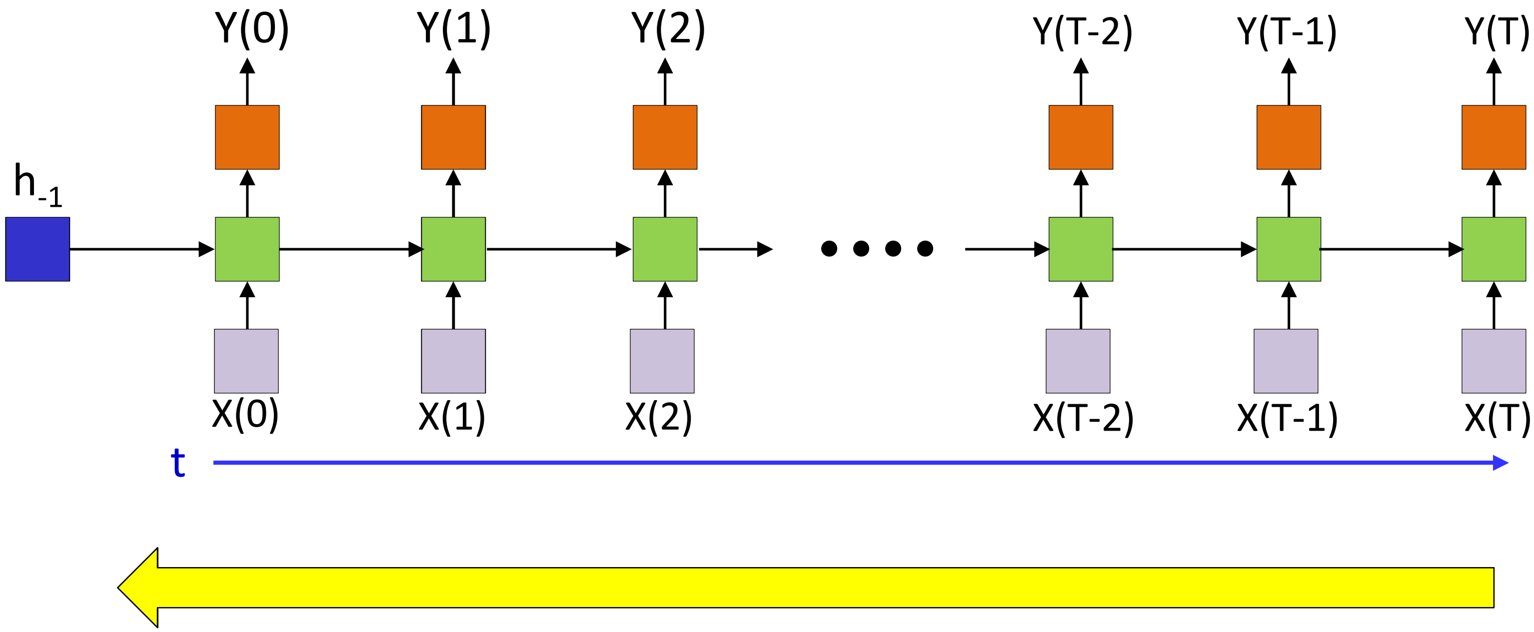
- Given a collection of *sequence* training instances comprising input sequences and output sequences of equal length, with one-to-one correspondence
  - $(\mathbf{X}_i, \mathbf{D}_i)$, where
  - $\mathbf{X}_i = X_{i,0}, \dots, X_{i,T}$
  - $\mathbf{D}_i = D_{i,0}, \dots, D_{i,T}$

Training: Forward pass



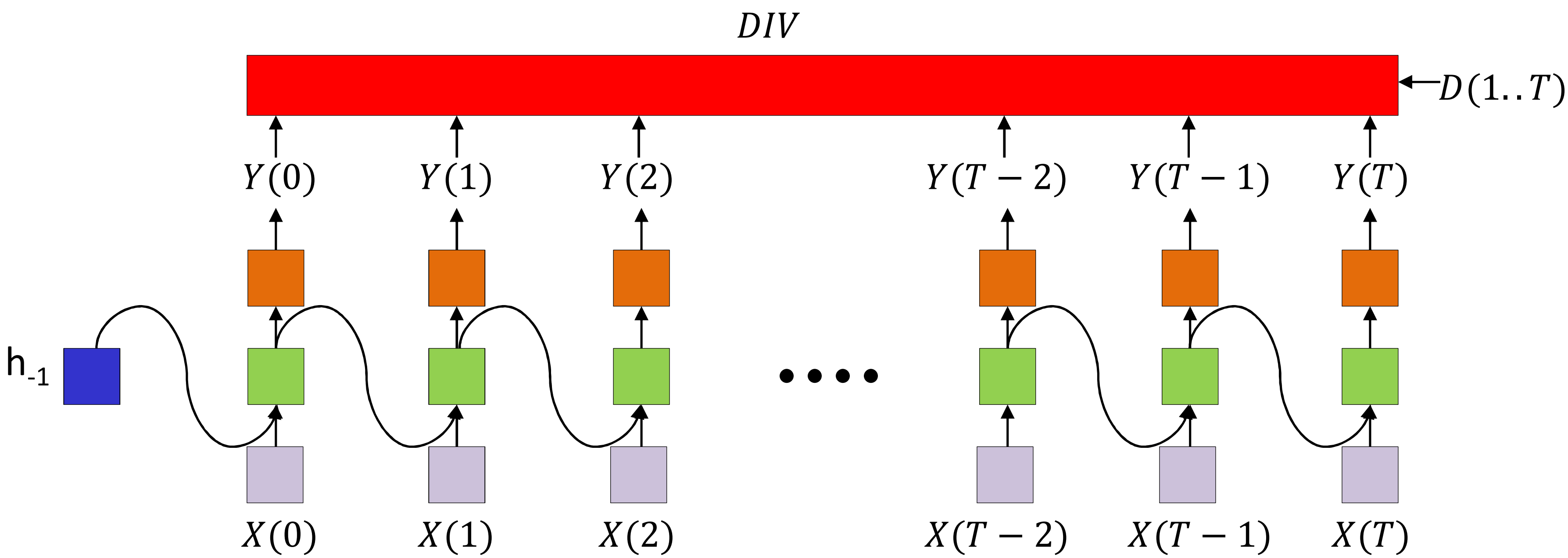
- For each training input:
- Forward pass: pass the entire data sequence through the network, generate outputs

Training: Computing gradients



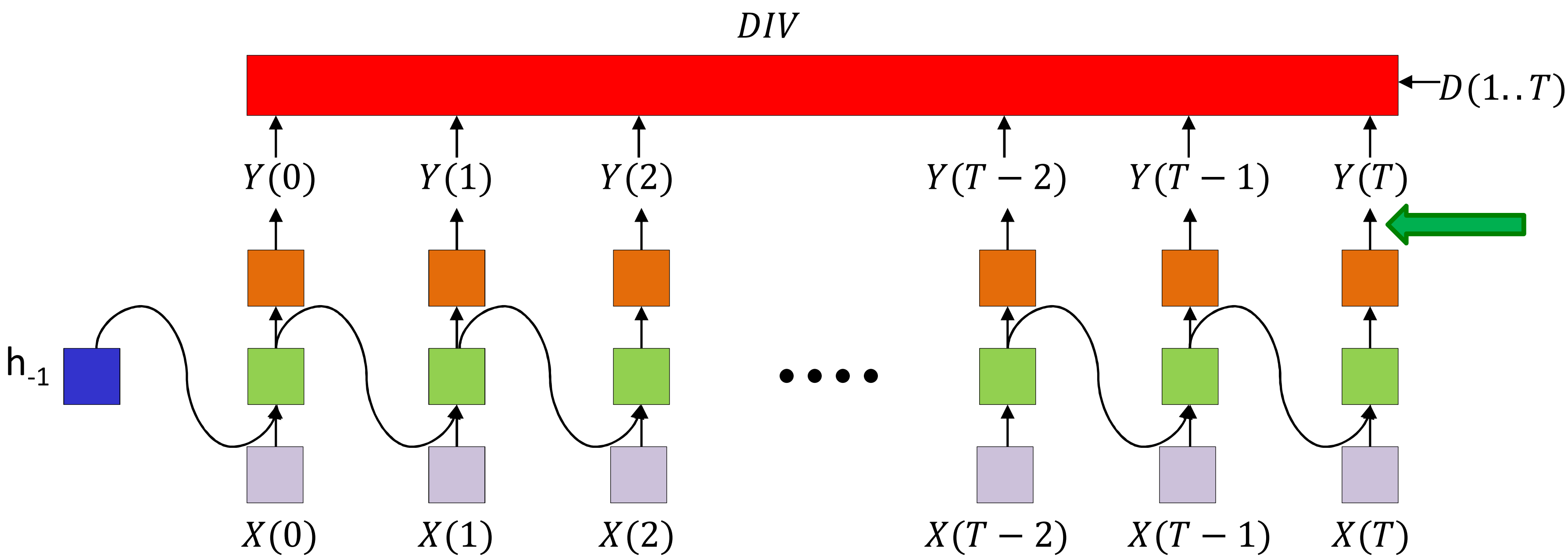
- For each training input:
- Backward pass: Compute gradients via backpropagation
  - *Back Propagation Through Time*

Back Propagation Through Time



- The divergence computed is between the *sequence of outputs* by the network and the *desired sequence of outputs*
- This is *not* just the sum of the divergences at individual times
  - Unless we explicitly define it that way

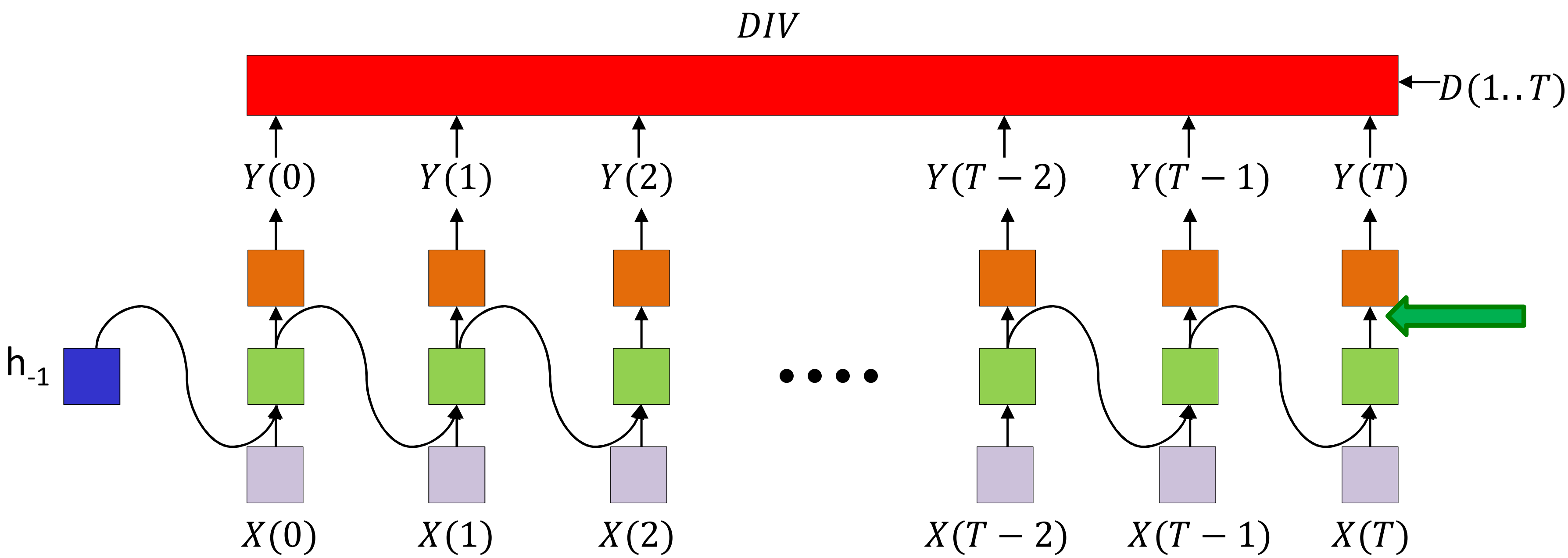
Back Propagation Through Time



First step of backprop: Compute $\nabla_{Y(t)} DIV$ for all t

The rest of backprop continues from there

Back Propagation Through Time

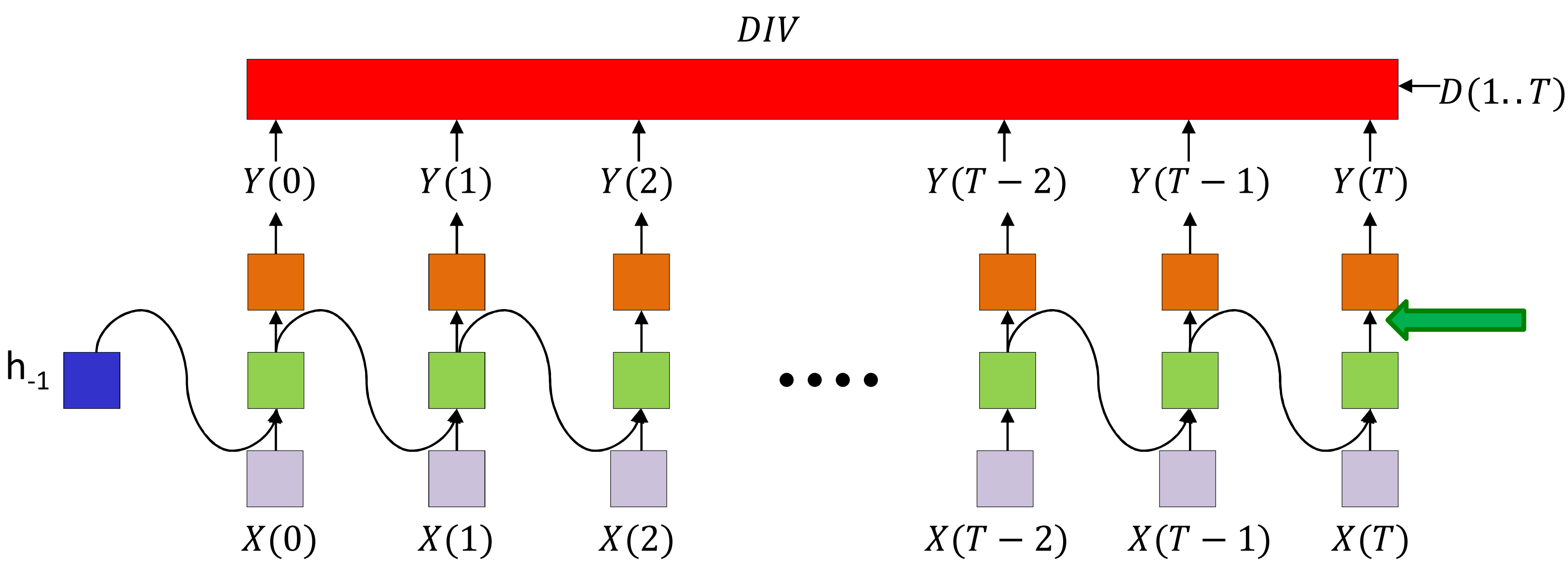


First step of backprop: Compute $\nabla_{Y(t)} DIV$ for all t

$$\nabla_{Z^{(1)}(t)} DIV = \nabla_{Y(t)} DIV \ \nabla_{Z(t)} Y(t)$$

And so on!

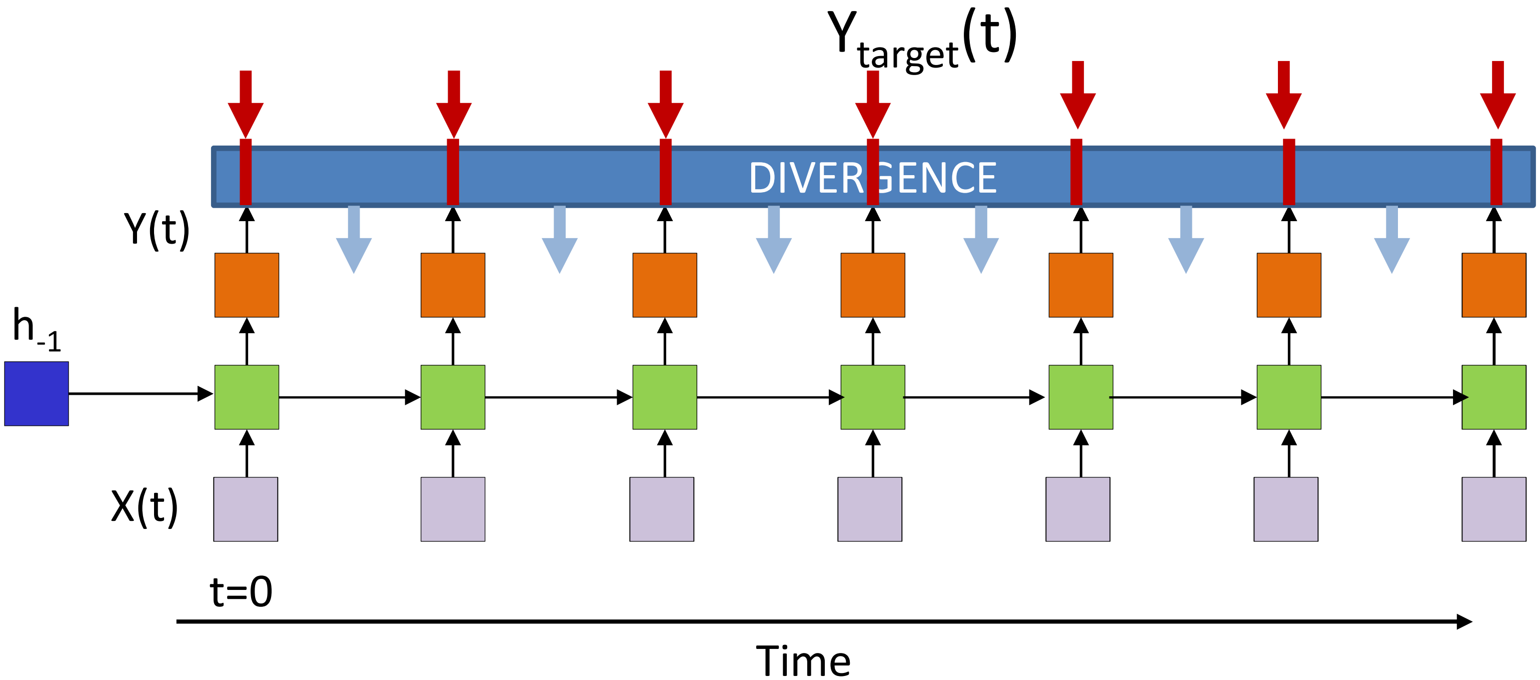
Back Propagation Through Time



First step of backprop: Compute $\nabla_{Y(t)}DIV$ for all t

- The key component is the computation of this derivative!!
- This depends on the definition of "DIV"

Time-synchronous recurrence

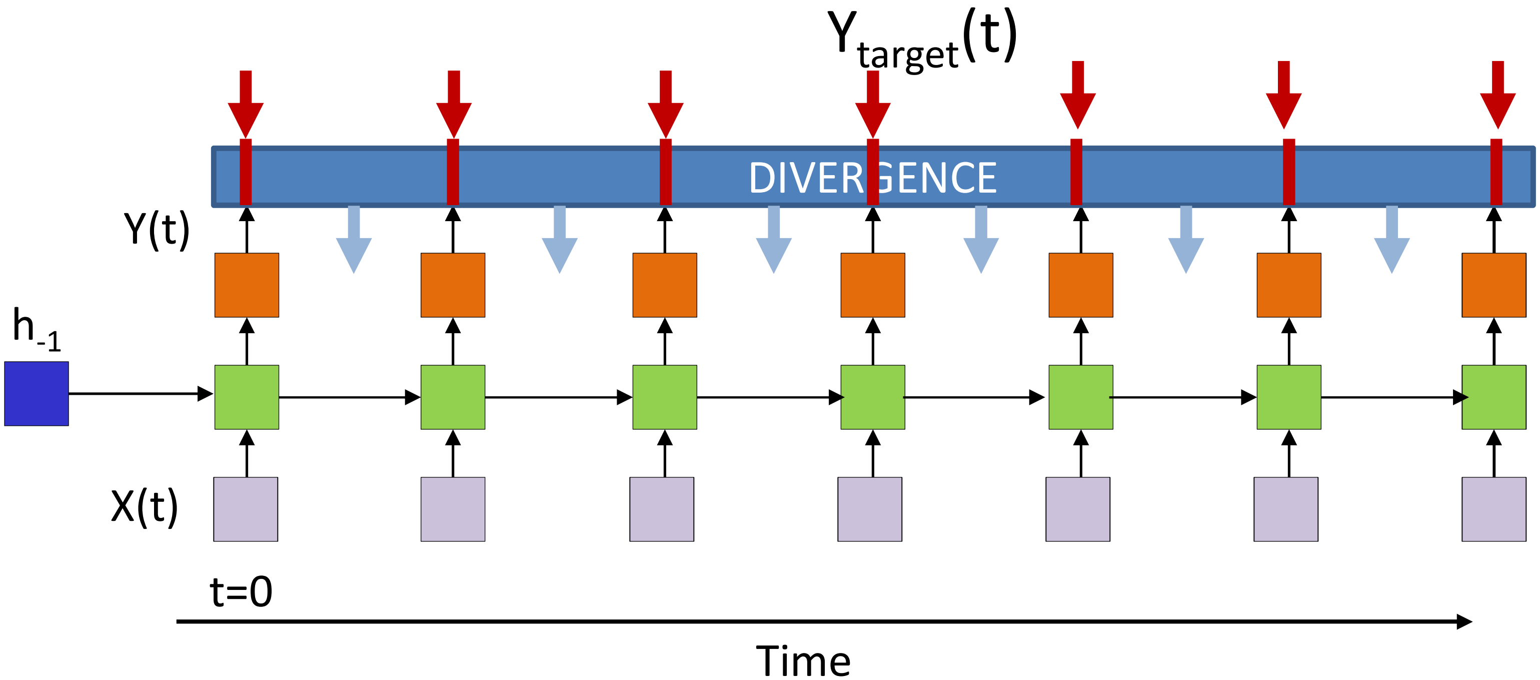


- Usual assumption: ***Sequence divergence is the sum of the divergence at individual instants***

$$Div\left(Y_{target}(1 \ldots T), Y(1 \ldots T)\right) = \sum_t Div\left(Y_{target}(t), Y(t)\right)$$

$$\nabla_{Y(t)} Div\left(Y_{target}(1 \ldots T), Y(1 \ldots T)\right) = \nabla_{Y(t)} Div\left(Y_{target}(t), Y(t)\right)$$

Time-synchronous recurrence



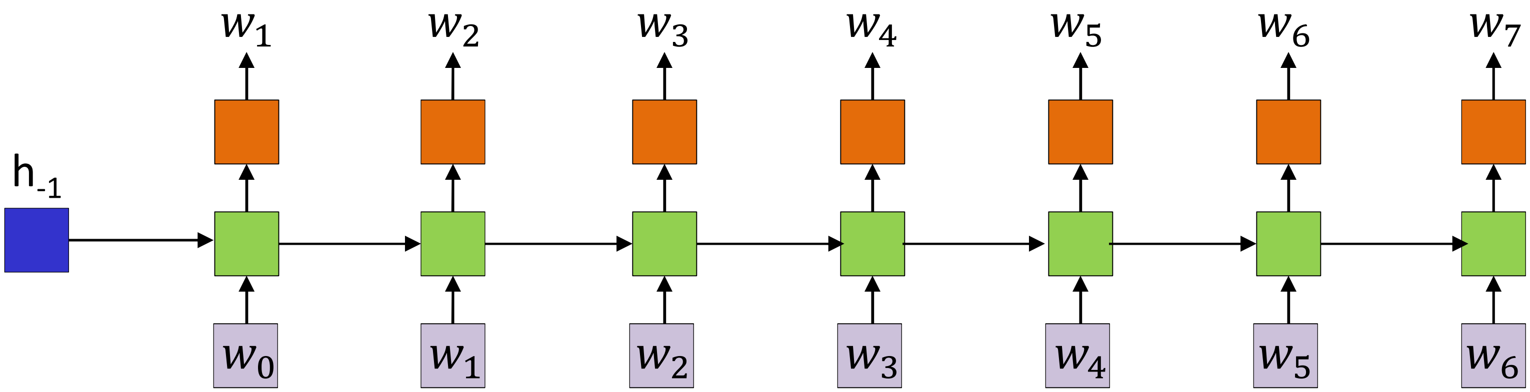
- Usual assumption: ***Sequence divergence is the sum of the divergence at individual instants***

$$Div\left(Y_{target}(1 \dots T), Y(1 \dots T)\right) = \sum_t Div\left(Y_{target}(t), Y(t)\right)$$

$$\nabla_{Y(t)} Div\left(Y_{target}(1 \dots T), Y(1 \dots T)\right) = \nabla_{Y(t)} Div\left(Y_{target}(t), Y(t)\right)$$

Typical Divergence for classification: $Div\left(Y_{target}(t), Y(t)\right) = Xent(Y_{target}, Y)$

Simple recurrence example: Text Modelling



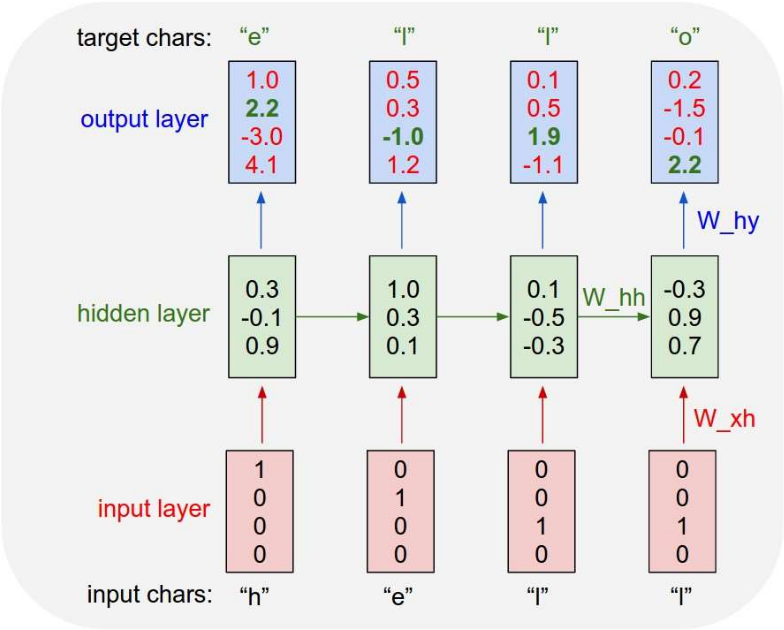
- Learn a model that can predict the next character given a sequence of characters
  - Or, at a higher level, words
- After observing inputs $w_0 \dots w_k$ it predicts w_{k+1}

Simple recurrence example: Text Modelling

Figure from Andrej Karpathy.

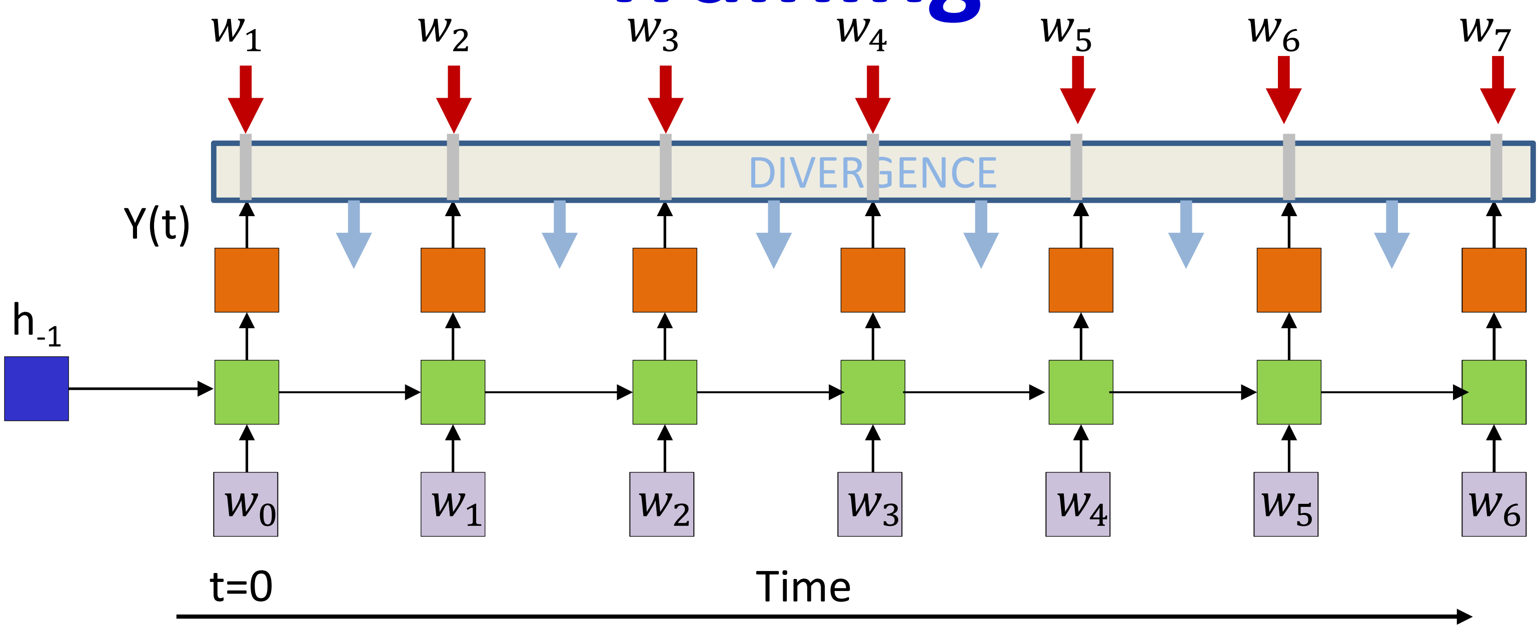
Input: Sequence of characters (presented as one-hot vectors).

Target output after observing "h e l l" is "o"



- Input presented as one-hot vectors
  - Actually "embeddings" of one-hot vectors
- Output: probability distribution over characters
  - Must ideally peak at the target character

Training



- Input: symbols as one-hot vectors
  - Dimensionality of the vector is the size of the "vocabulary"
- Output: Probability distribution over symbols

$$Y(t,i) = P(V_i | w_0 \dots w_{t-1})$$

  - V_i is the i-th symbol in the vocabulary
- Divergence

The probability assigned to the correct next word

$$Div\left(Y_{target}(1 \dots T), Y(1 \dots T)\right) = \sum_t Xent\left(Y_{target}(t), Y(t)\right) = -\sum_t \log Y(t, w_{t+1})$$

Brief detour: Language models

- Modelling language using time-synchronous nets

- More generally language models and embeddings..

Which open source project?

```
/*
 * 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_SECON
  return segtable;
}
```

Language modelling using RNNs

Four score and seven years ???

A B R A H A M L I N C O L ??

- Problem: Given a sequence of words (or characters) predict the next one

Language modelling: Representing words

- Represent words as one-hot vectors
  - Pre-specify a vocabulary of N words in fixed (e.g. lexical) order
    - *E.g. [A AARDVARK AARON ABACK ABACUS… ZZYP]*
  - Represent each word by an N-dimensional vector with N-1 zeros and a single 1 (in the position of the word in the ordered list of words)
    - E.g. "AARDVARK" → [0 1 0 0 0 …]
    - E.g. "AARON" → [0 0 1 0 0 0 …]

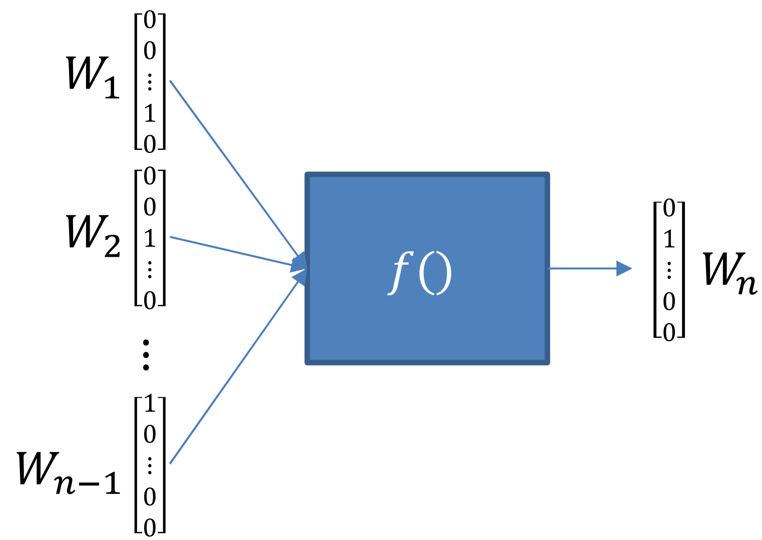
- Characters can be similarly represented
  - English will require about 100 characters, to include both cases, special characters such as commas, hyphens, apostrophes, etc., and the space character

Predicting words

Four score and seven years ???

$$W_n = f(W_1, \dots, W_{n-1})$$

Nx1 one-hot vectors



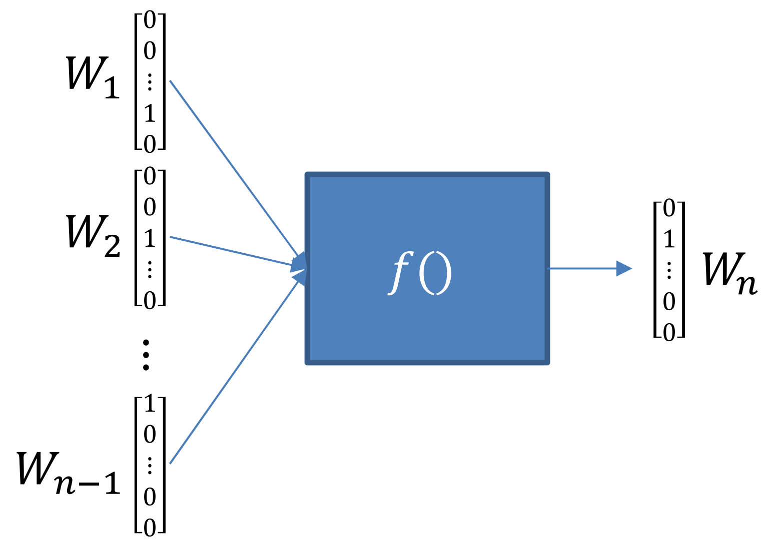
- Given one-hot representations of $W_1 \dots W_{n-1}$, predict W_n

Predicting words

Four score and seven years ???

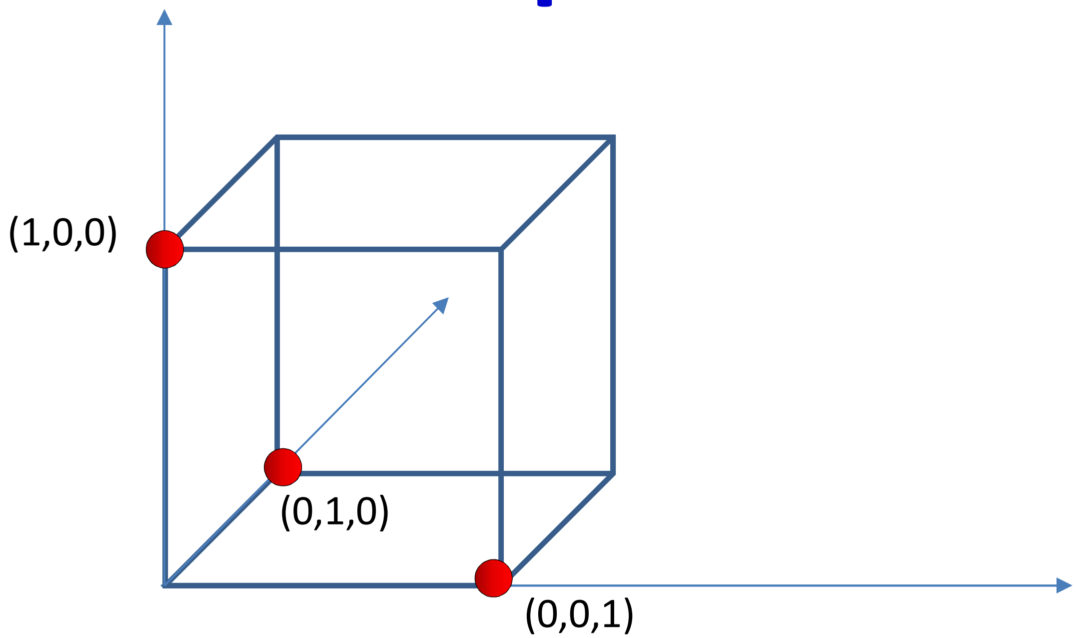
$$W_n = f(W_1, \ldots, W_{n-1})$$

Nx1 one-hot vectors



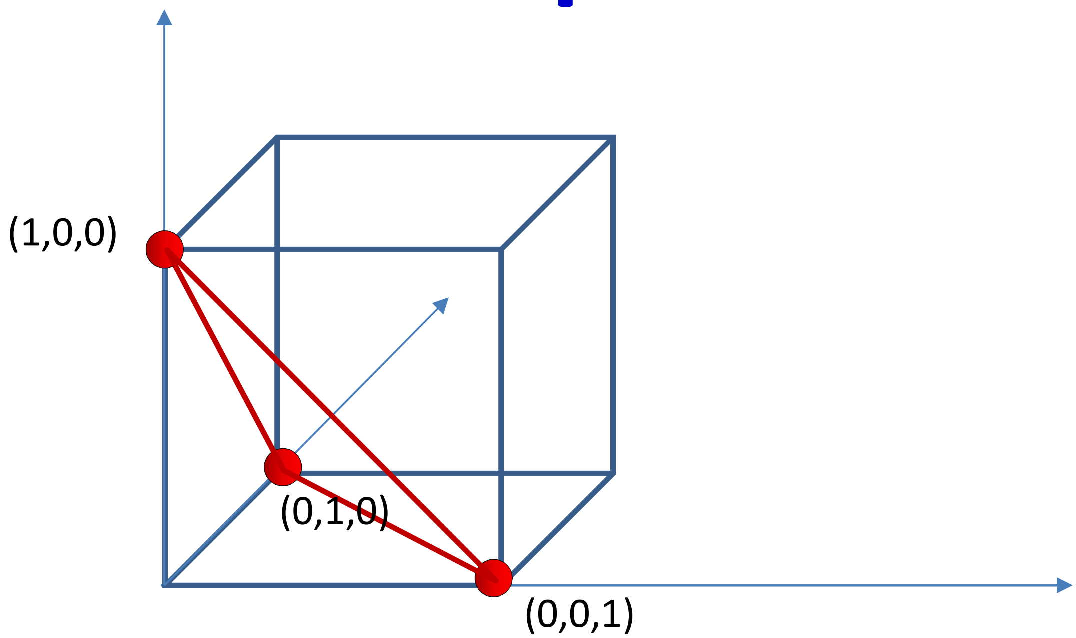
- Given one-hot representations of $W_1 \ldots W_{n-1}$, predict W_n
- **Dimensionality problem:** All inputs $W_1 \ldots W_{n-1}$ are both very high-dimensional and very sparse

The one-hot representation



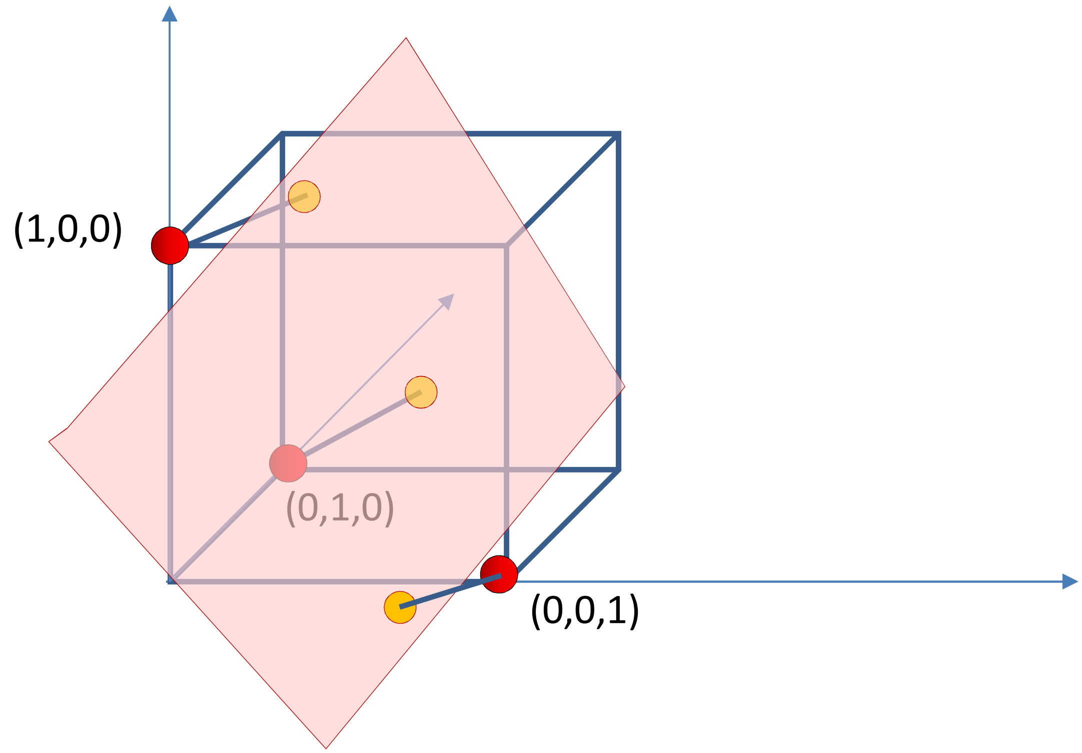
- The one hot representation uses only N corners of the 2^N corners of a unit cube
  - Actual volume of space used = 0
    - $(1, \varepsilon, \delta)$ has no meaning except for $\varepsilon = \delta = 0$
  - Density of points: $\mathcal{O}\left(\frac{N}{r^N}\right)$
- This is a tremendously inefficient use of dimensions

Why one-hot representation



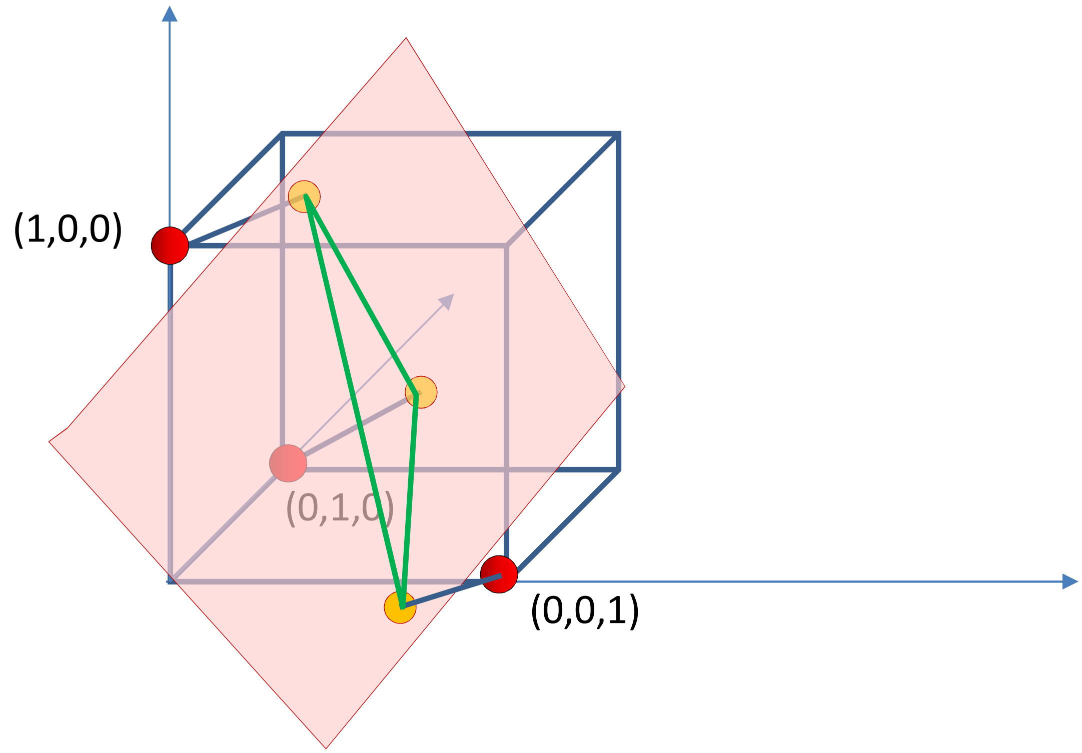
- The one-hot representation makes no assumptions about the relative importance of words
  – All word vectors are the same length
- It makes no assumptions about the relationships between words
  – The distance between every pair of words is the same

Solution to dimensionality problem



- Project the points onto a lower-dimensional subspace
  - The volume used is still 0, but density can go up by many orders of magnitude
    - Density of points: $\mathcal{O}\left(\frac{N}{r^M}\right)$

Solution to dimensionality problem

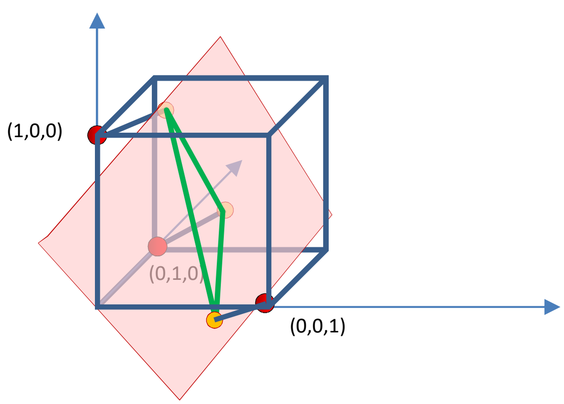


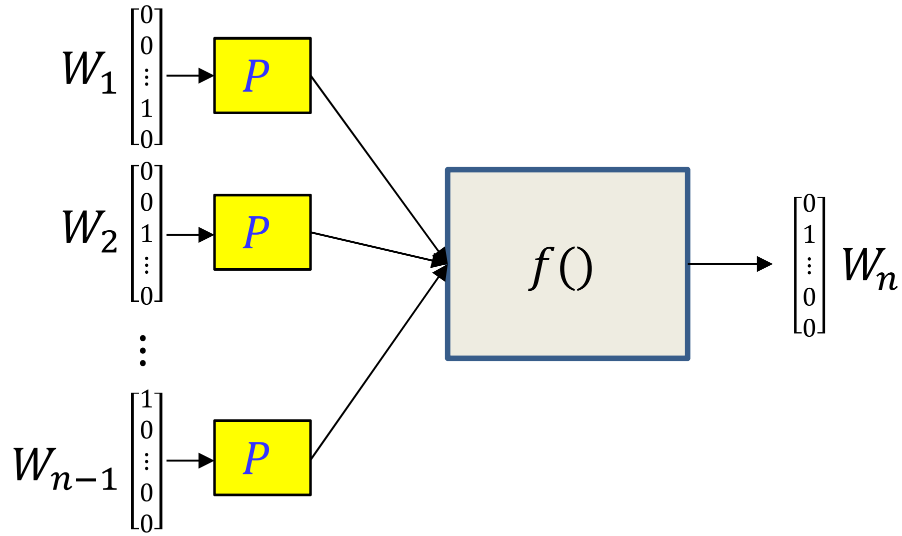
- Project the points onto a lower-dimensional subspace
  - The volume used is still 0, but density can go up by many orders of magnitude
    - Density of points: $\mathcal{O}\left(\frac{N}{r^M}\right)$
  - If properly learned, the distances between projected points will capture semantic relations between the words
    - This will also require linear transformation (stretching/shrinking/rotation) of the subspace

The *Projected* word vectors

Four score and seven years ???

$$W_n = f(PW_1, PW_2, \ldots, PW_{n-1})$$

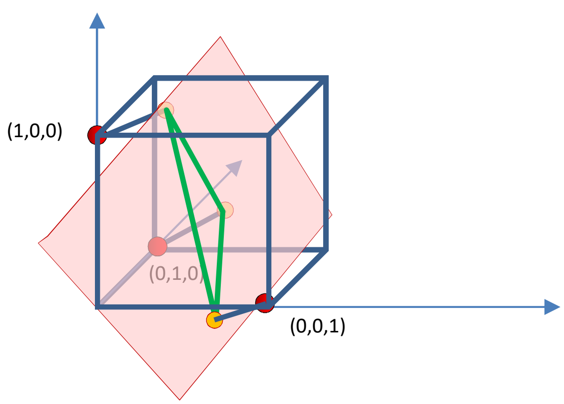


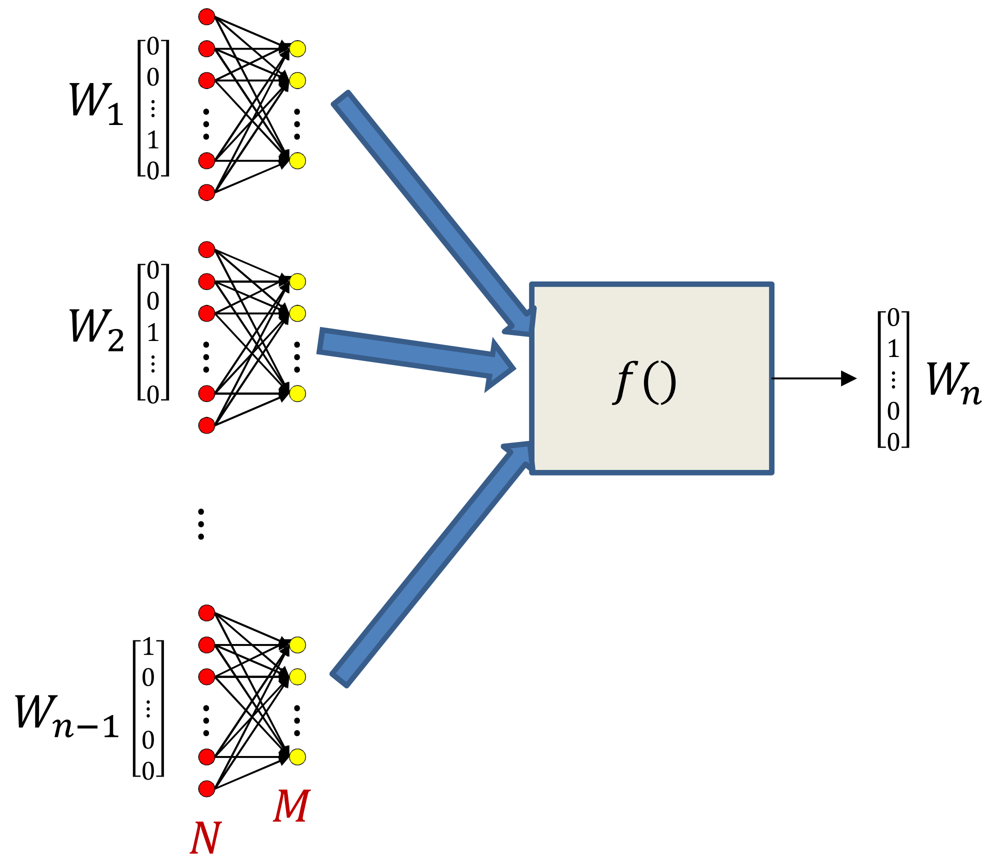


- *Project* the N-dimensional one-hot word vectors into a lower-dimensional space
  - Replace every one-hot vector W_i by PW_i
  - P is an $M \times N$ matrix
  - PW_i is now an M-dimensional vector
  - *Learn P* using an appropriate objective
    - Distances in the projected space will reflect relationships imposed by the objective

"Projection"

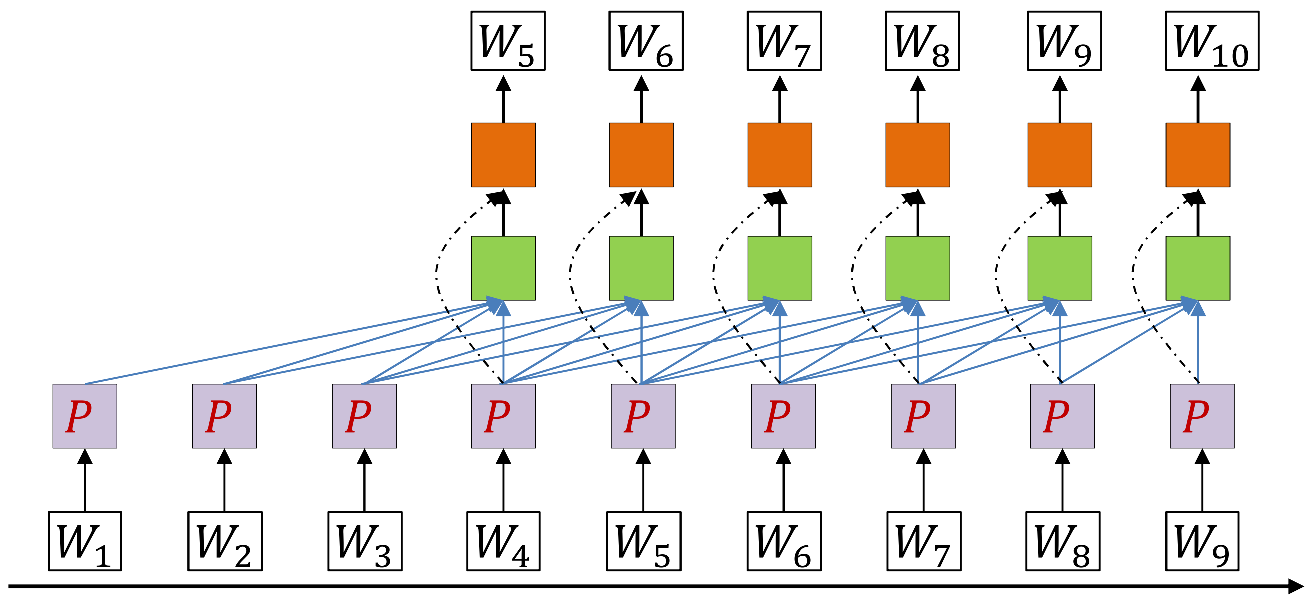
$$W_n = f(PW_1, PW_2, \ldots, PW_{n-1})$$





- P is a simple linear transform
- A single transform can be implemented as a layer of M neurons with linear activation
- The transforms that apply to the individual inputs are all M-neuron linear-activation subnets with tied weights

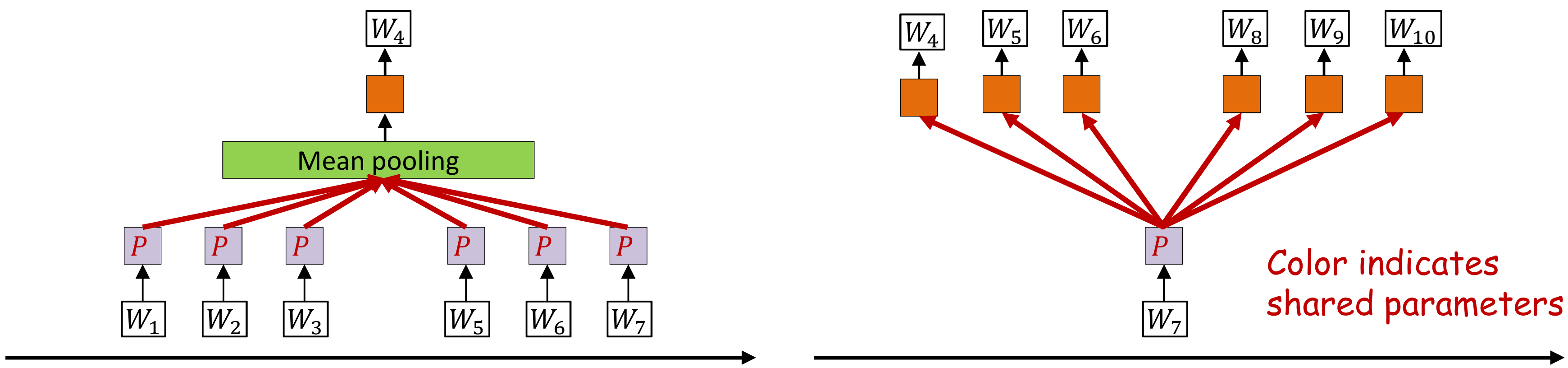
Predicting words: The TDNN model



- Predict each word based on the past N words
  - "A neural probabilistic language model", Bengio et al. 2003
  - Hidden layer has Tanh() activation, output is softmax

- One of the outcomes of learning this model is that we also learn low-dimensional representations PW of words

Alternative models to learn projections



- Soft bag of words: Predict word based on words in immediate context
  - Without considering specific position
- Skip-grams: Predict adjacent words based on current word
- More on these in a future recitation?

Embeddings: Examples

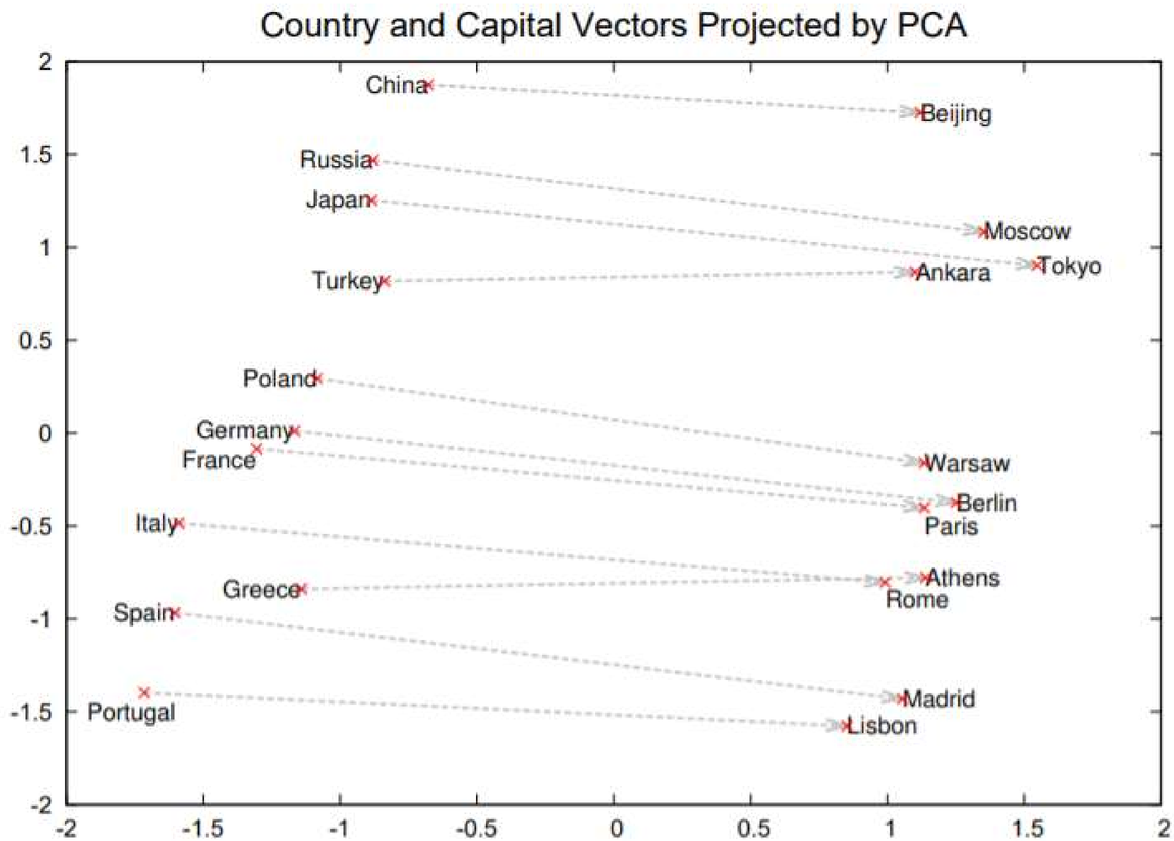
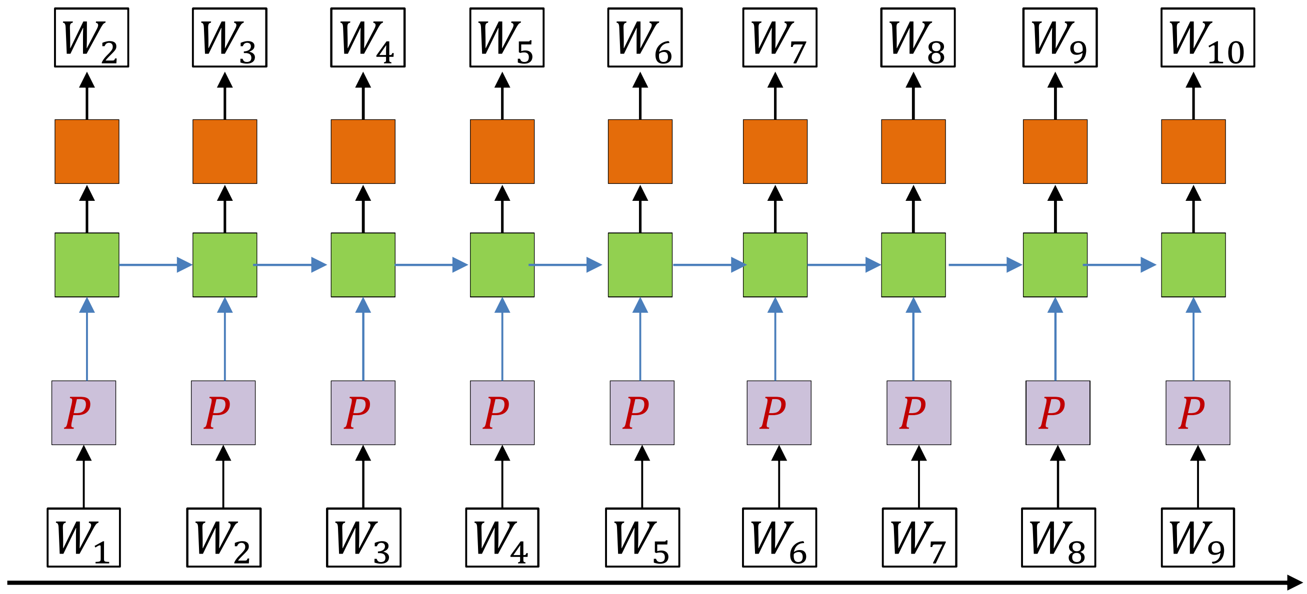


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

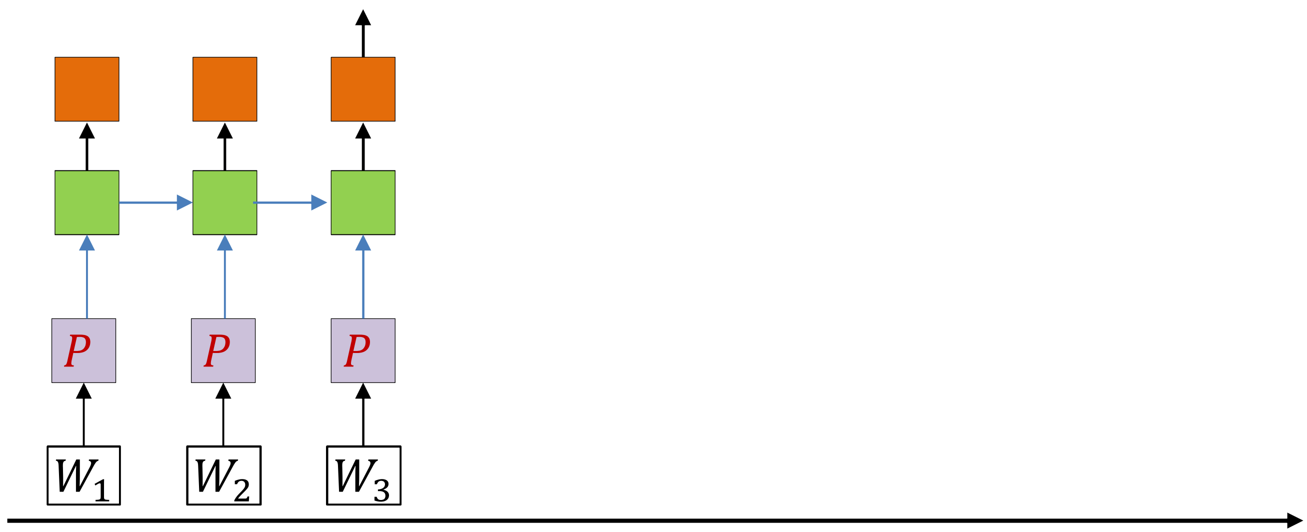
- From Mikolov et al., 2013, "Distributed Representations of Words and Phrases and their Compositionality"

Generating Language: The model



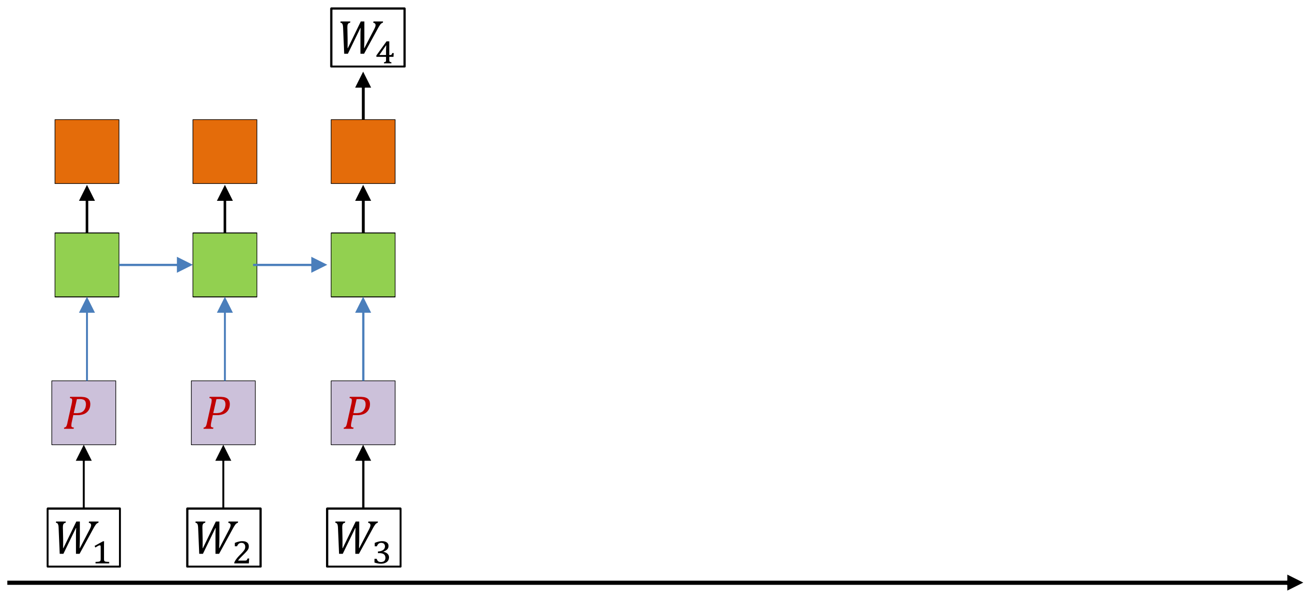
- The hidden units are (one or more layers of) LSTM units
- Trained via backpropagation from a lot of text

Generating Language: Synthesis



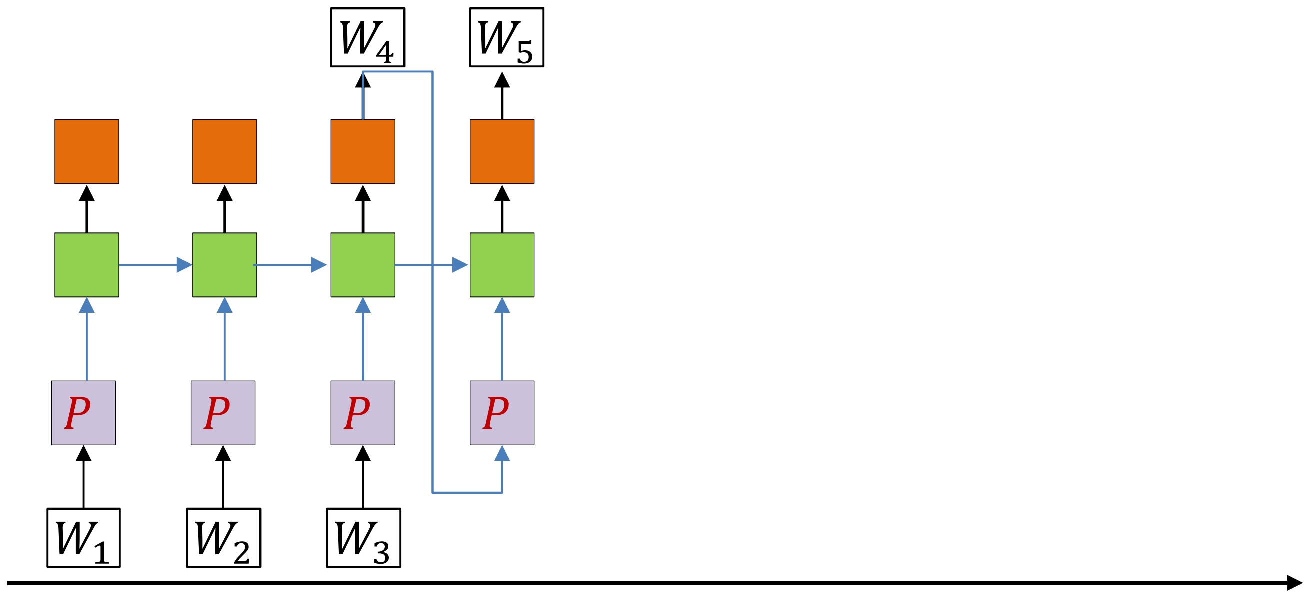
- On trained model : Provide the first few words
  - One-hot vectors
- After the last input word, the network generates a probability distribution over words
  - Outputs an N-valued probability distribution rather than a one-hot vector

Generating Language: Synthesis



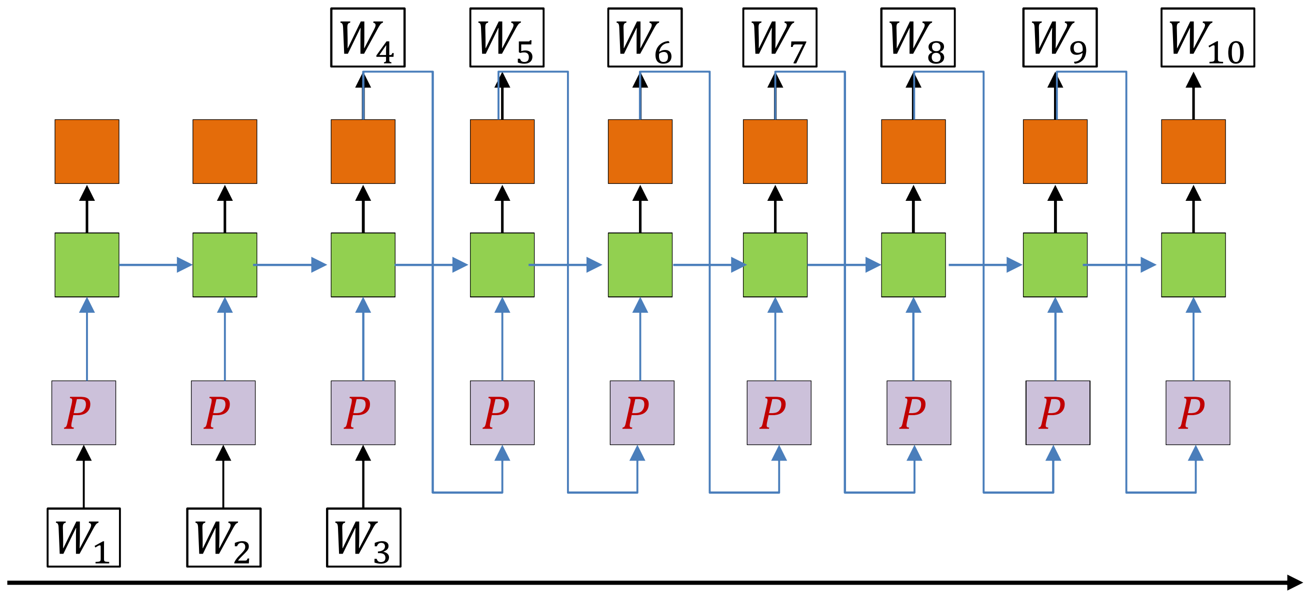
- On trained model : Provide the first few words
  - One-hot vectors
- After the last input word, the network generates a probability distribution over words
  - Outputs an N-valued probability distribution rather than a one-hot vector
- Draw a word from the distribution
  - And set it as the next word in the series

Generating Language: Synthesis



- Feed the drawn word as the next word in the series
  - And draw the next word from the output probability distribution

Generating Language: Synthesis



- Feed the drawn word as the next word in the series
  - And draw the next word from the output probability distribution
- Continue this process until we terminate generation
  - In some cases, e.g. generating programs, there may be a natural termination

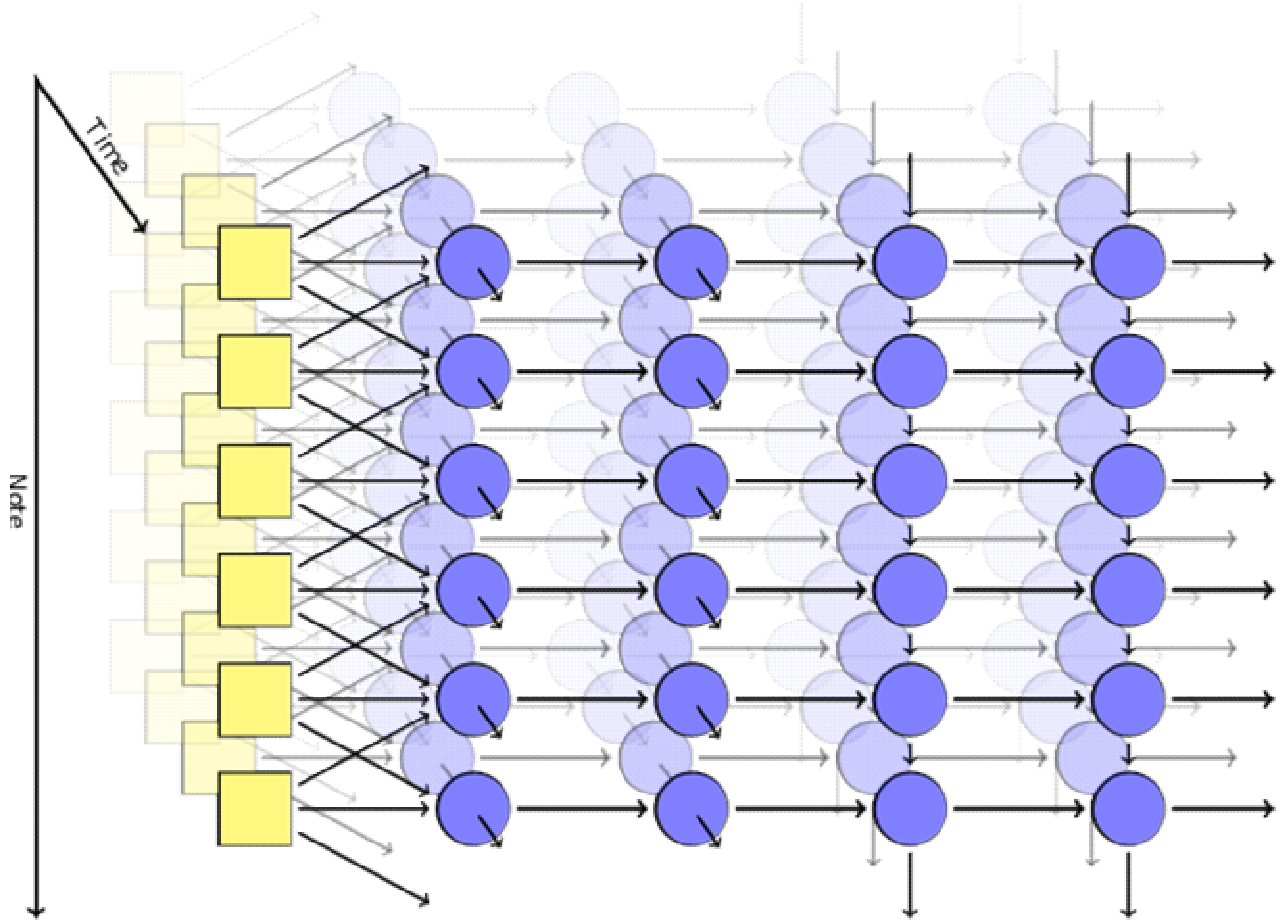
Which open source project?

```
/*
 * 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_SECON
  return segtable;
}
```

Trained on linux source code

Actually uses a *character-level* model (predicts character sequences)

Composing music with RNN

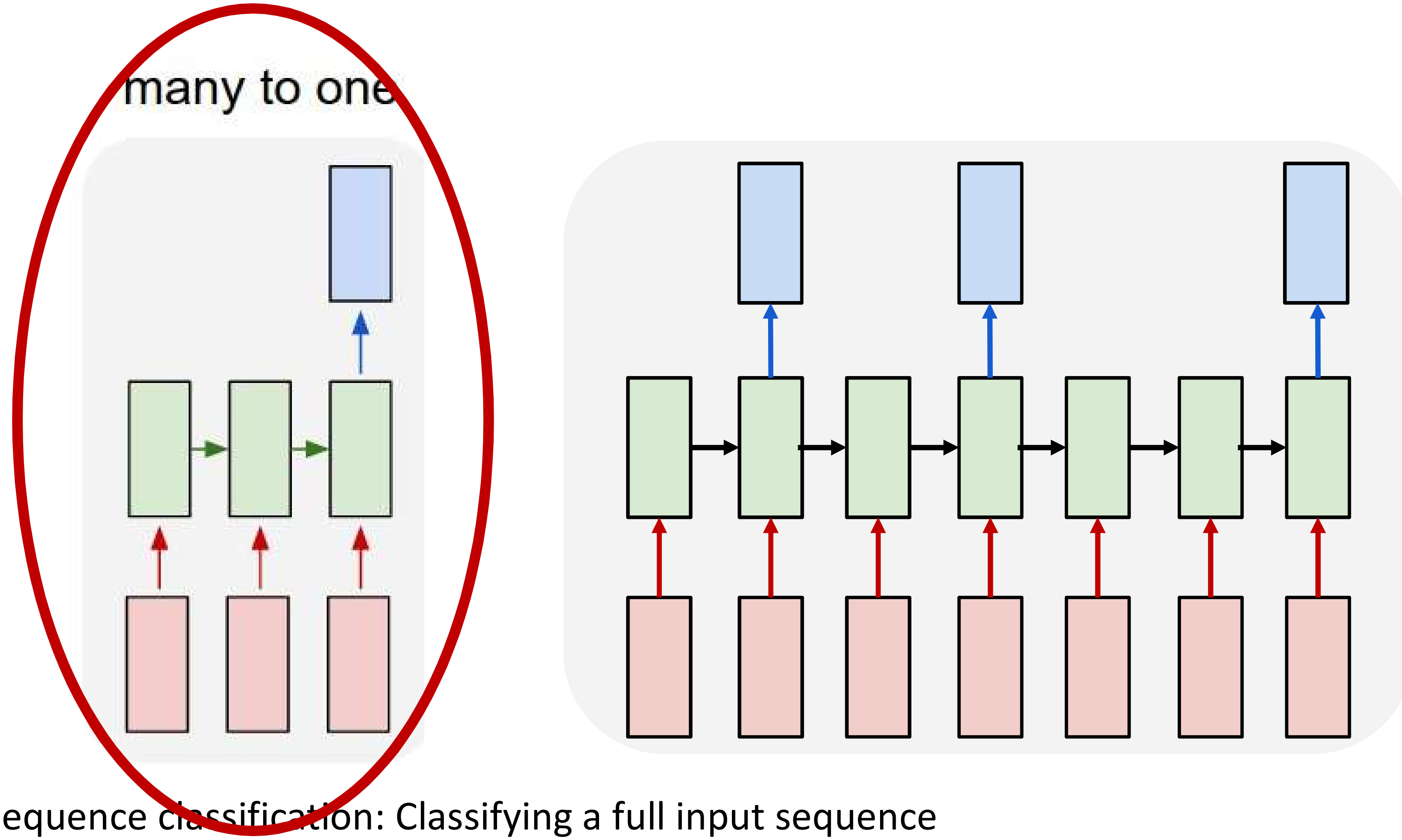


http://www.hexahedria.com/2015/08/03/composing-music-with-recurrent-neural-networks/

Returning to our problem

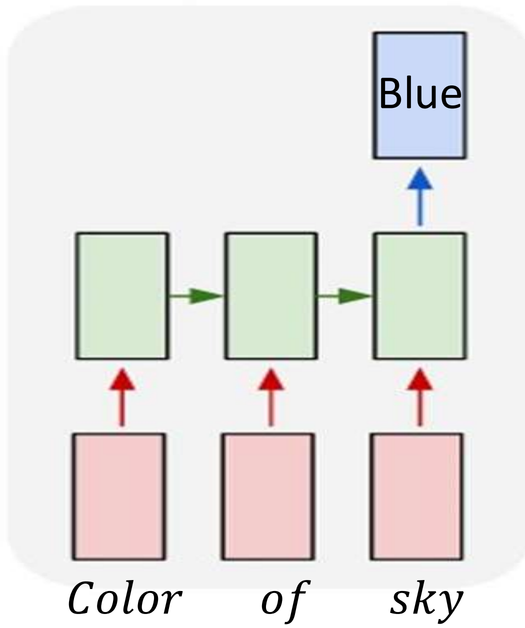
- Divergences are harder to define in other scenarios..

Variants on recurrent nets



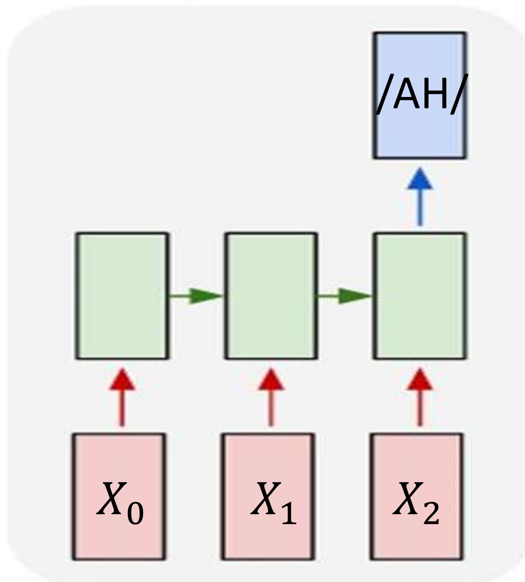
- Sequence classification: Classifying a full input sequence
  - E.g phoneme recognition
- Order synchronous , time asynchronous sequence-to-sequence generation
  - E.g. speech recognition
  - Exact location of output is unknown a priori

Example..



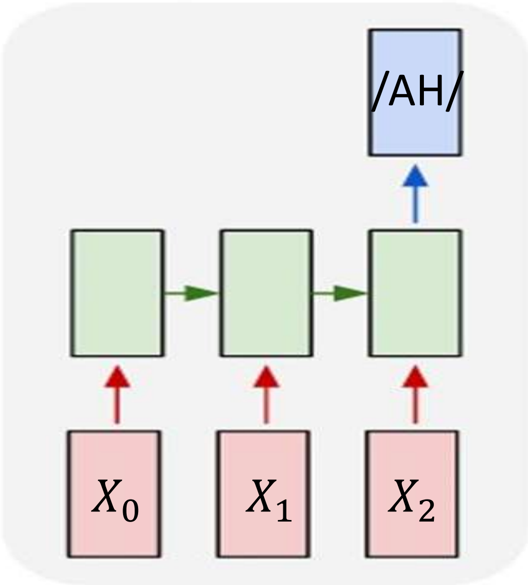
- Question answering
- Input : Sequence of words
- Output: Answer at the end of the question

Example..



- Speech recognition
- Input : Sequence of feature vectors (e.g. Mel spectra)
- Output: Phoneme ID at the end of the sequence
  - Represented as an N-dimensional output probability vector, where N is the number of phonemes

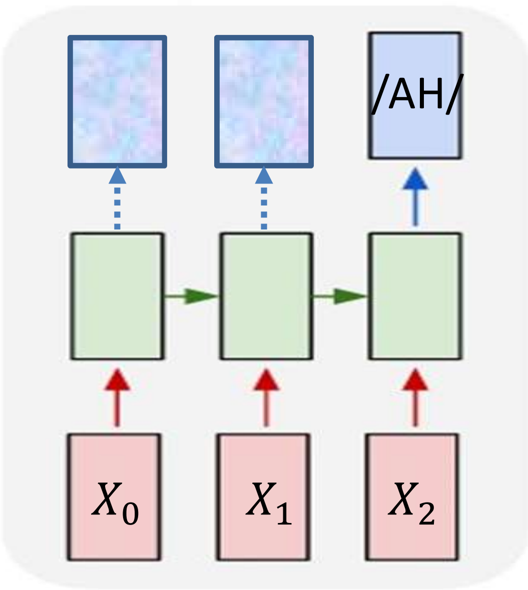
Inference: Forward pass



- Exact input sequence provided
  - Output generated when the last vector is processed
    - Output is a probability distribution over phonemes

- But what about at *intermediate stages?*

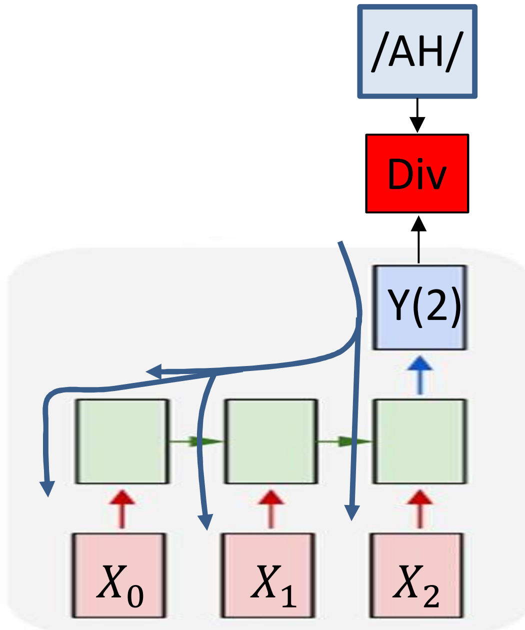
Forward pass



- Exact input sequence provided
  - Output generated when the last vector is processed
    - Output is a probability distribution over phonemes

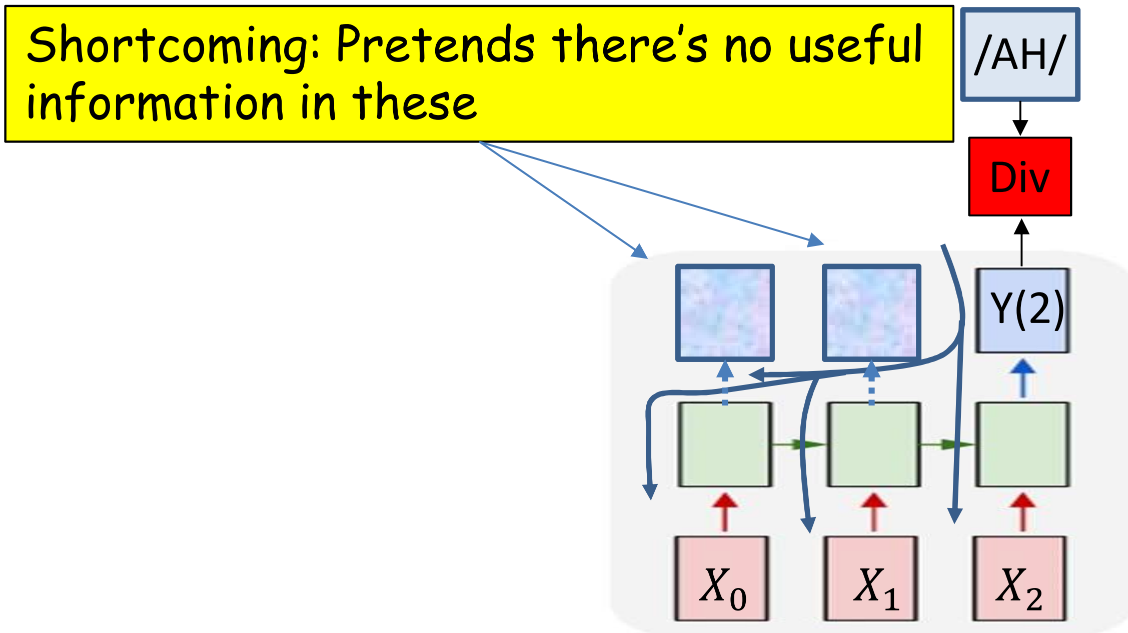
- Outputs are actually produced for *every* input
  - We only *read* it at the end of the sequence

Training



- The Divergence is only defined at the final input
  - $DIV\left(Y_{target}, Y\right) = Xent(Y(T), Phoneme)$
- This divergence must propagate through the net to update all parameters

Training

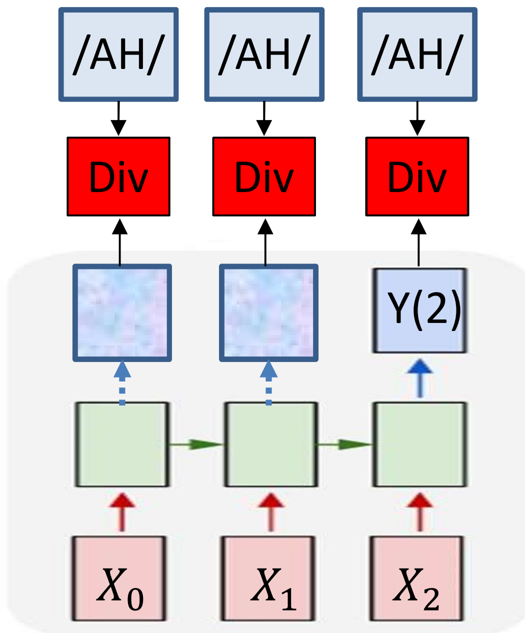


- The Divergence is only defined at the final input
  - $DIV\left(Y_{target}, Y\right) = Xent(Y(T), Phoneme)$
- This divergence must propagate through the net to update all parameters

Training

Fix: Use these outputs too.

These too must ideally point to the correct phoneme



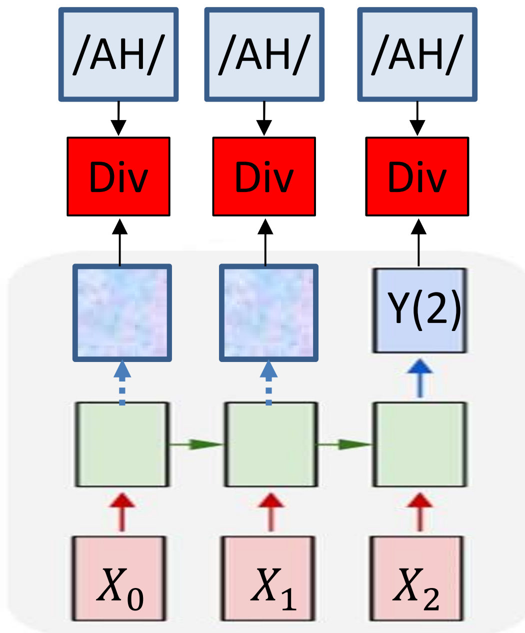
- Exploiting the untagged inputs: assume the same output for the entire input
- Define the divergence everywhere

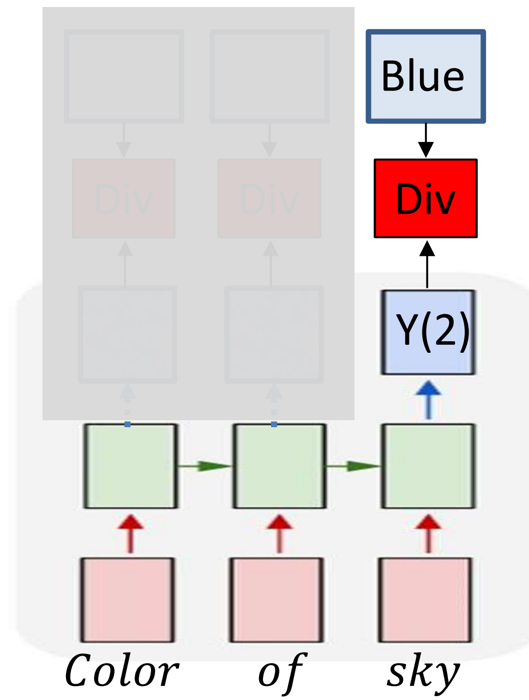
$$DIV\left(Y_{target}, Y\right) = \sum_t w_t Xent(Y(t), Phoneme)$$

Training

Fix: Use these outputs too.

These too must ideally point to the correct phoneme



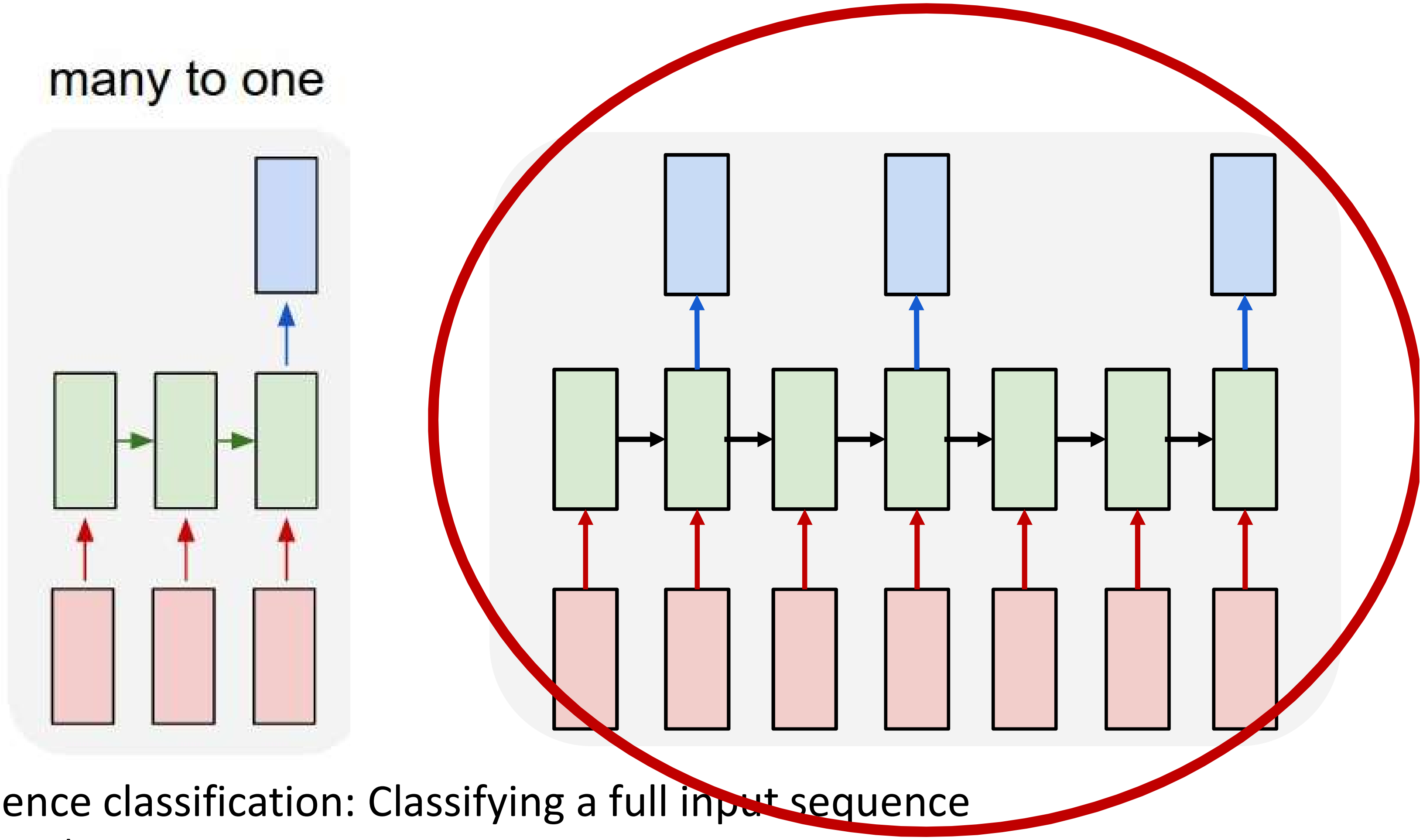


- Define the divergence everywhere

$$DIV\left(Y_{target}, Y\right) = \sum_t w_t Xent(Y(t), Phoneme)$$

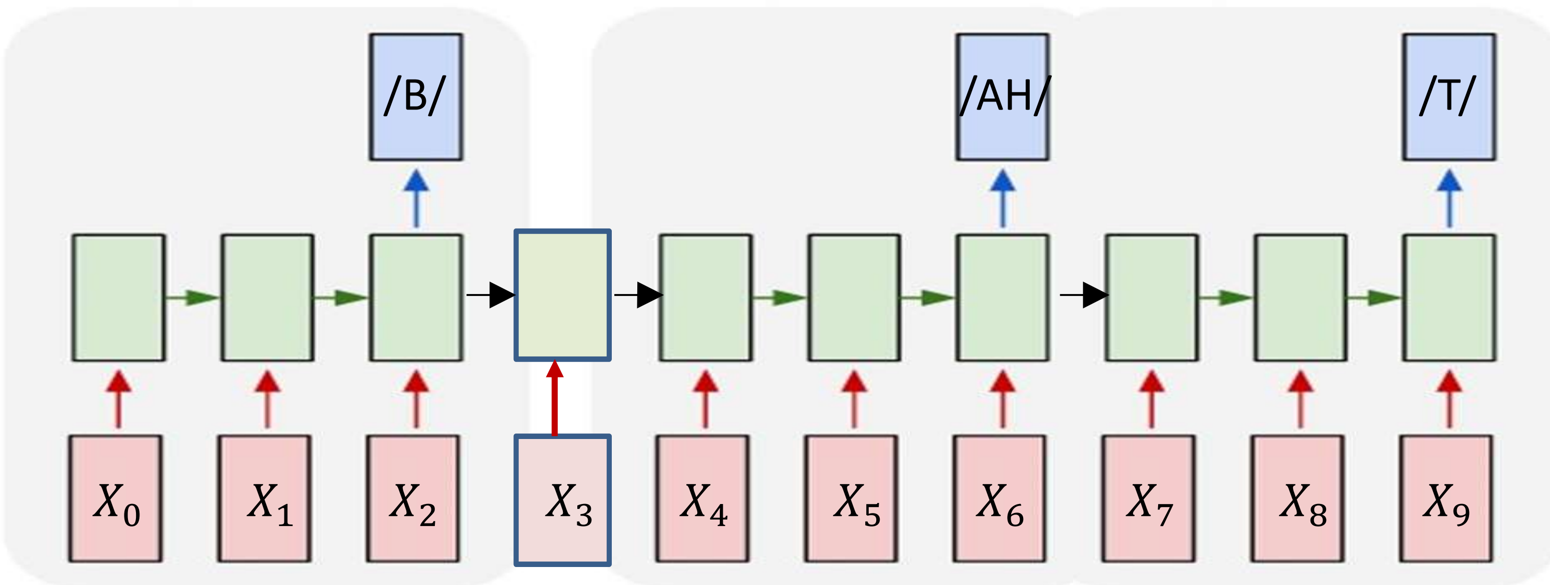
- Typical weighting scheme for speech: all are equally important
- Problem like question answering: answer only expected after the question ends
  - Only w_T is high, other weights are 0 or low

Variants on recurrent nets



- Sequence classification: Classifying a full input sequence
  - E.g phoneme recognition
- Order synchronous , time asynchronous sequence-to-sequence generation
  - E.g. speech recognition
  - Exact location of output is unknown a priori

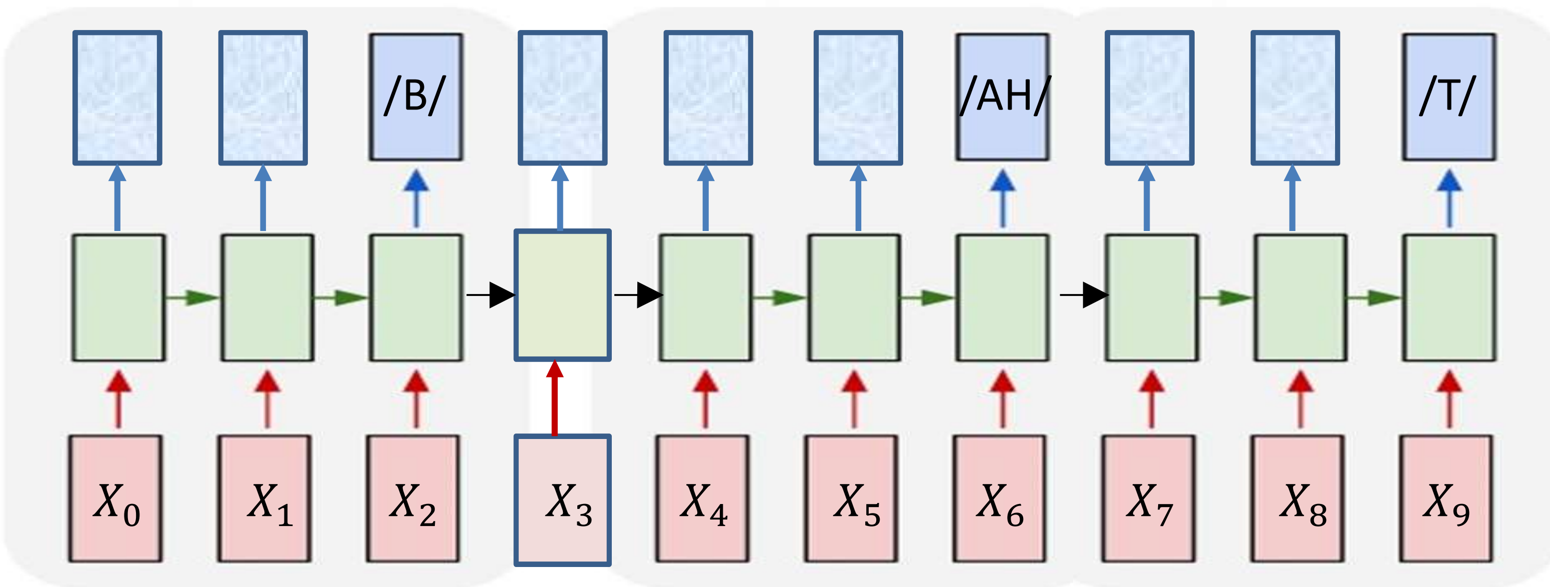
A more complex problem



- Objective: Given a sequence of inputs, asynchronously output a sequence of symbols
  – This is just a simple concatenation of many copies of the simple "output at the end of the input sequence" model we just saw

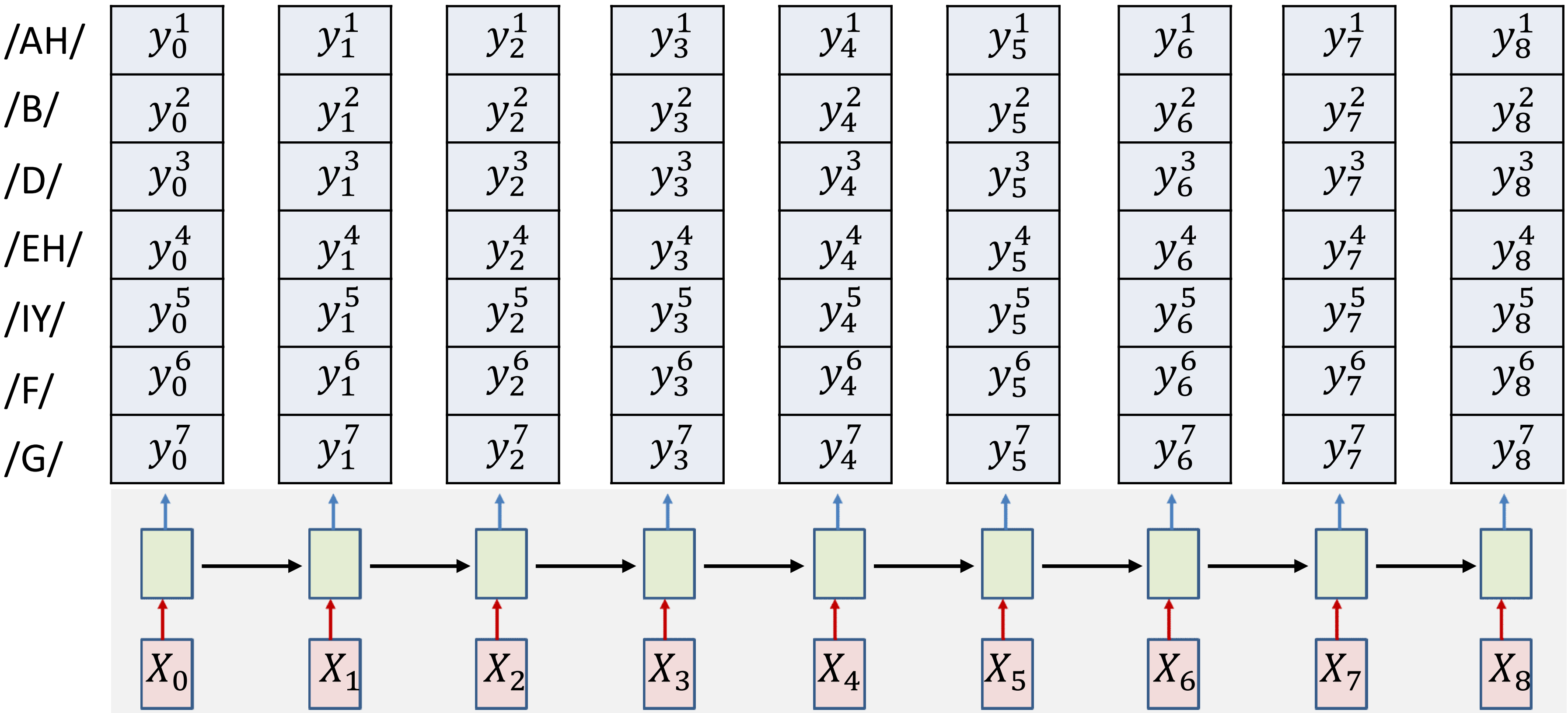
- But this simple extension complicates matters..

The *sequence-to-sequence* problem



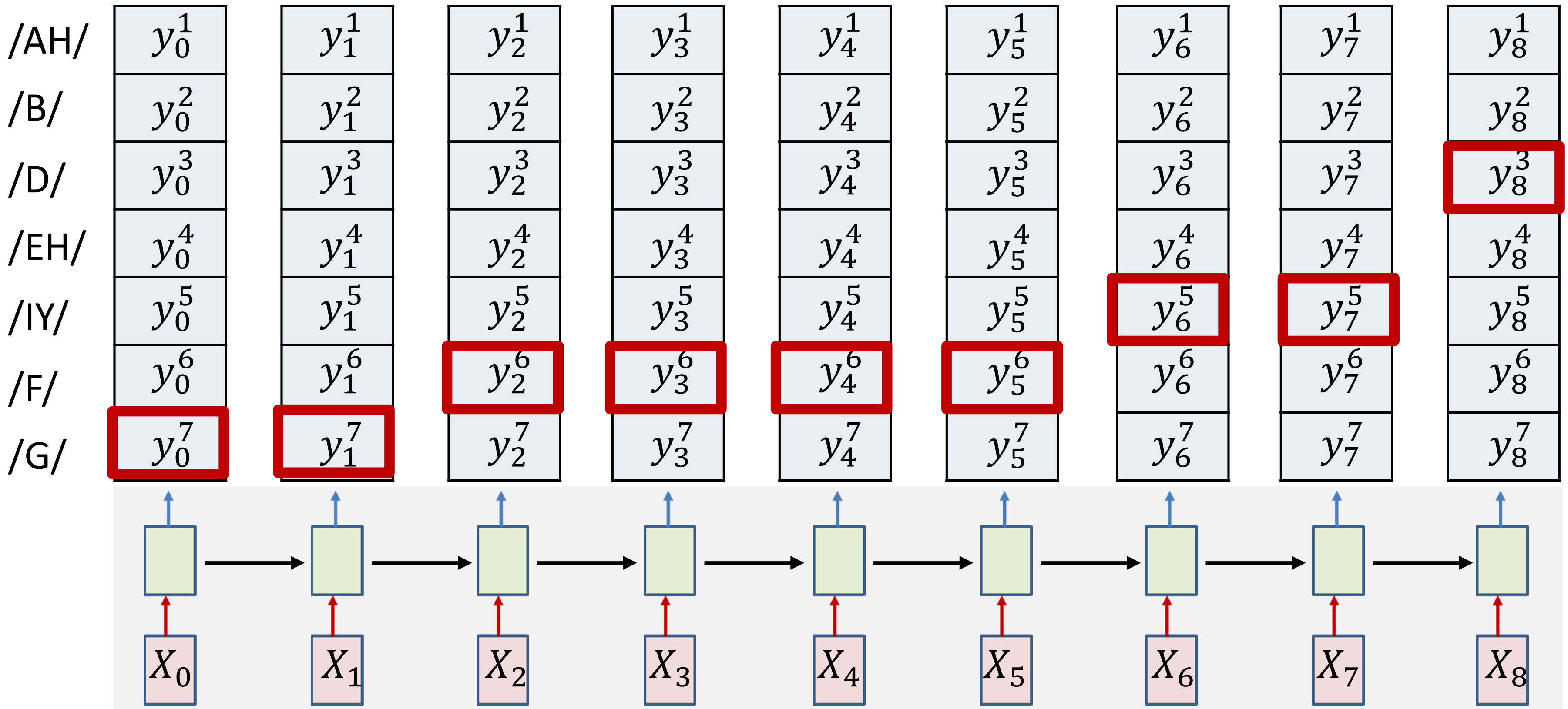
- How do we know *when* to output symbols
  - In fact, the network produces outputs at *every* time
  - *Which* of these are the *real* outputs
    - Outputs that represent the definitive occurrence of a symbol

The actual output of the network



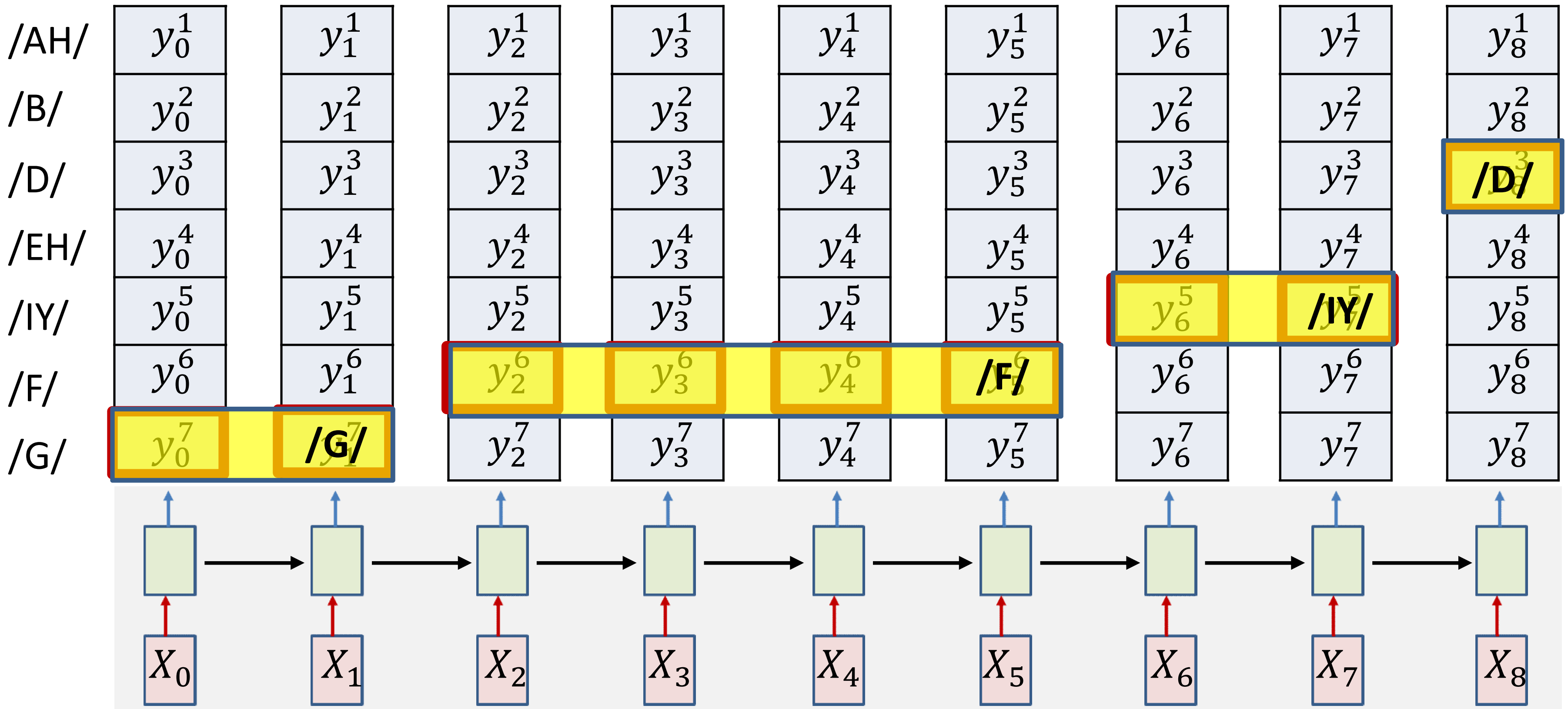
- At each time the network outputs a probability for *each* output symbol

The actual output of the network



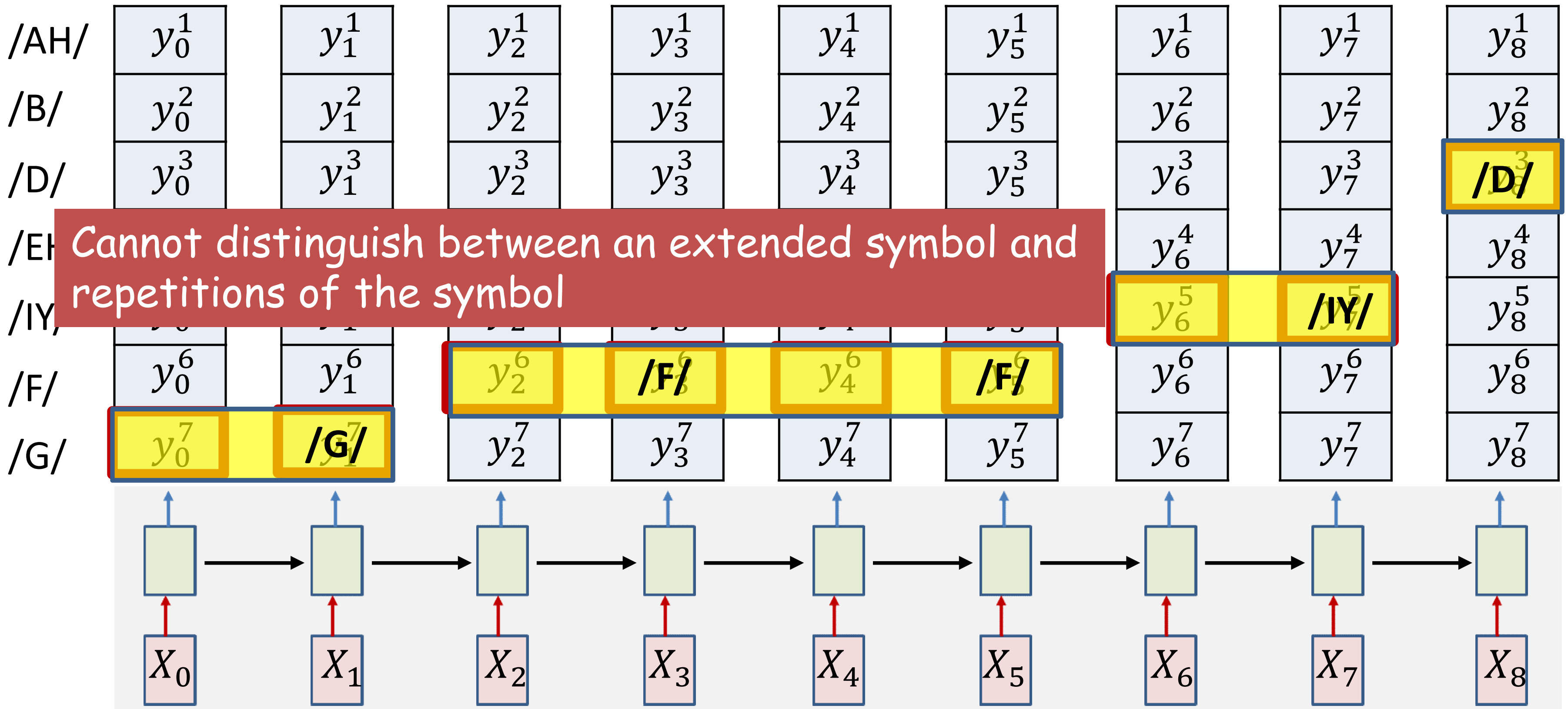
- Option 1: Simply select the most probable symbol at each time

The actual output of the network



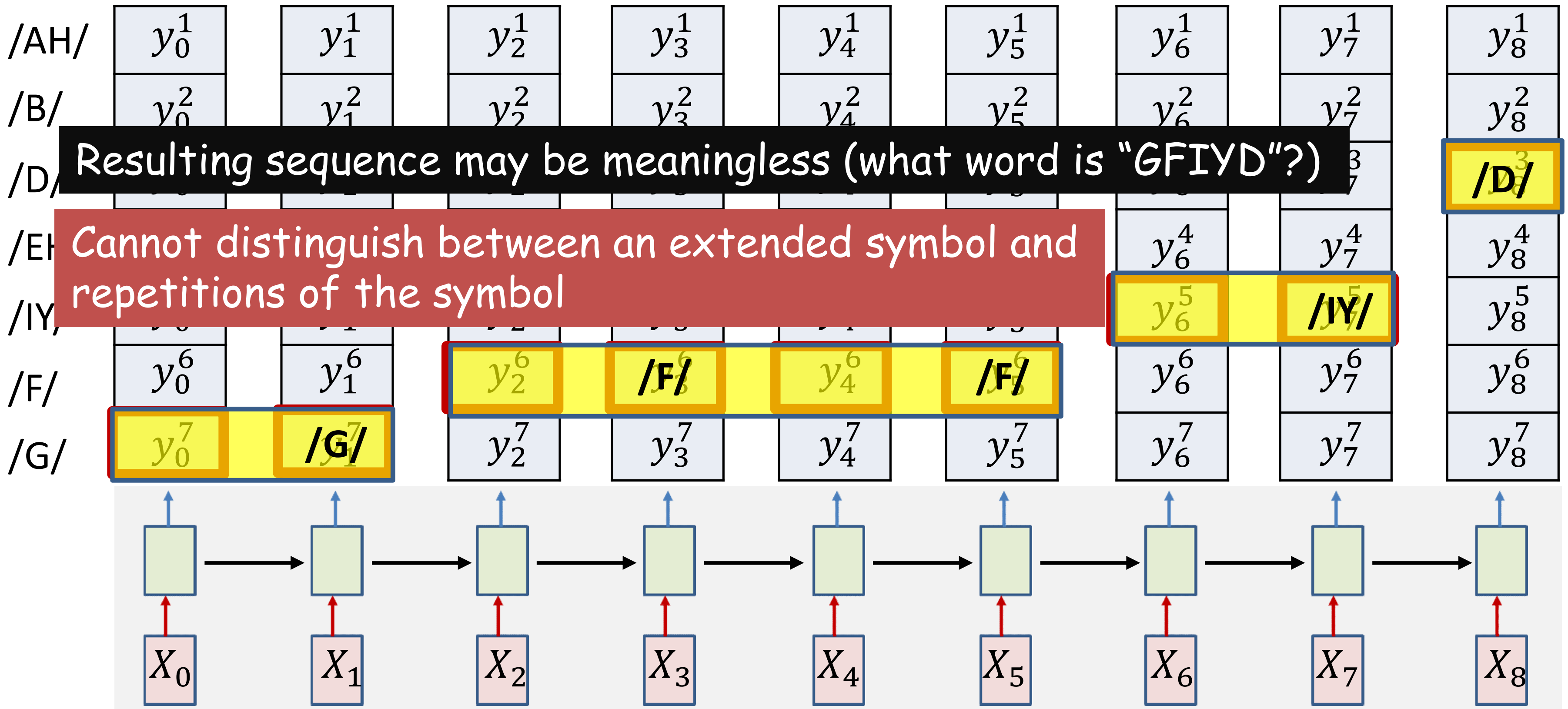
- Option 1: Simply select the most probable symbol at each time
  - *Merge* adjacent repeated symbols, and place the actual emission of the symbol in the final instant

The actual output of the network



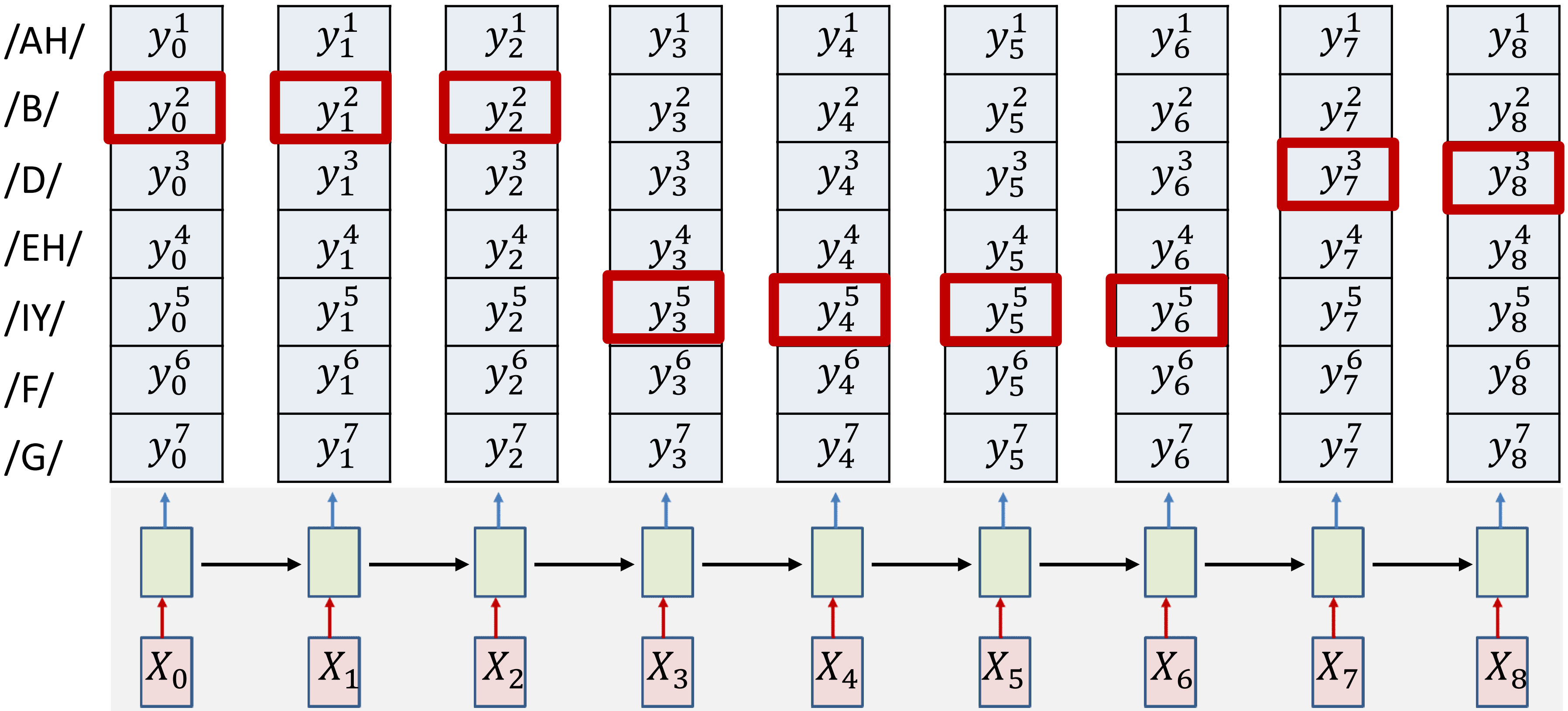
- Option 1: Simply select the most probable symbol at each time
  - *Merge* adjacent repeated symbols, and place the actual emission of the symbol in the final instant

The actual output of the network



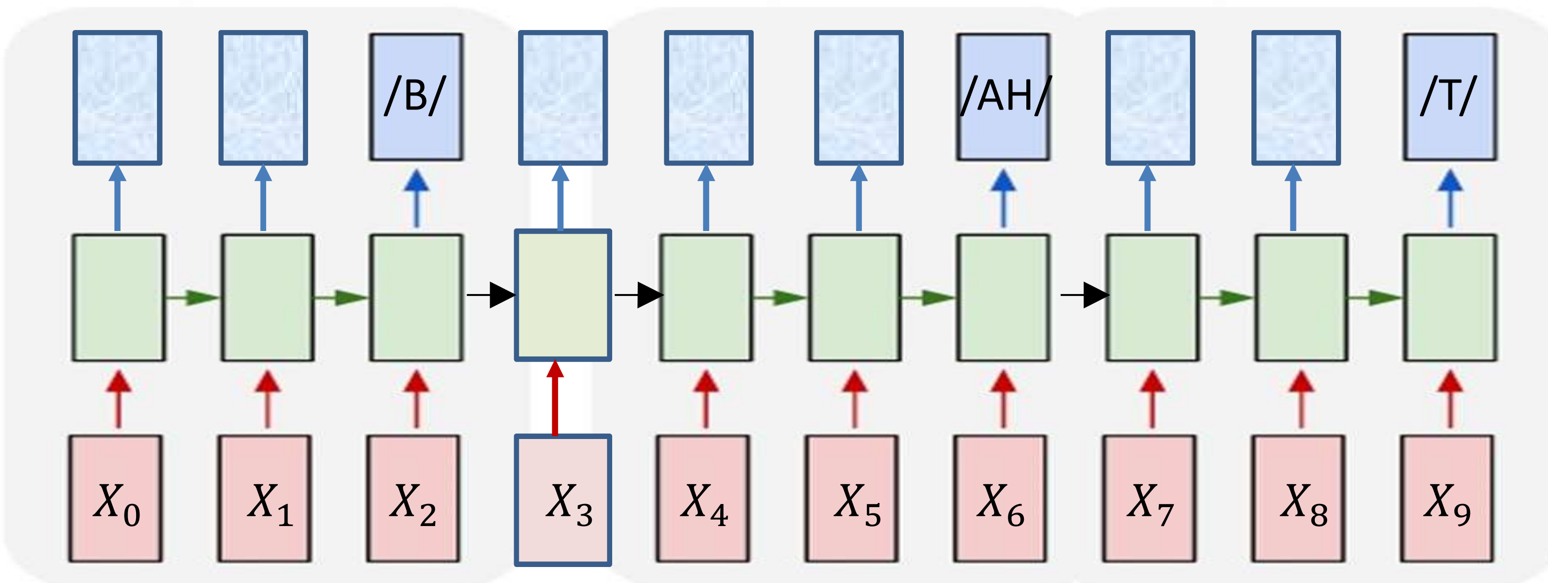
- Option 1: Simply select the most probable symbol at each time
  - *Merge* adjacent repeated symbols, and place the actual emission of the symbol in the final instant

The actual output of the network



- Option 2: Impose external constraints on what sequences are allowed
  - *E.g.* only allow sequences corresponding to dictionary words
  - *E.g. Sub-symbol* units (like in HW1 – what were they?)

The *sequence-to-sequence* problem

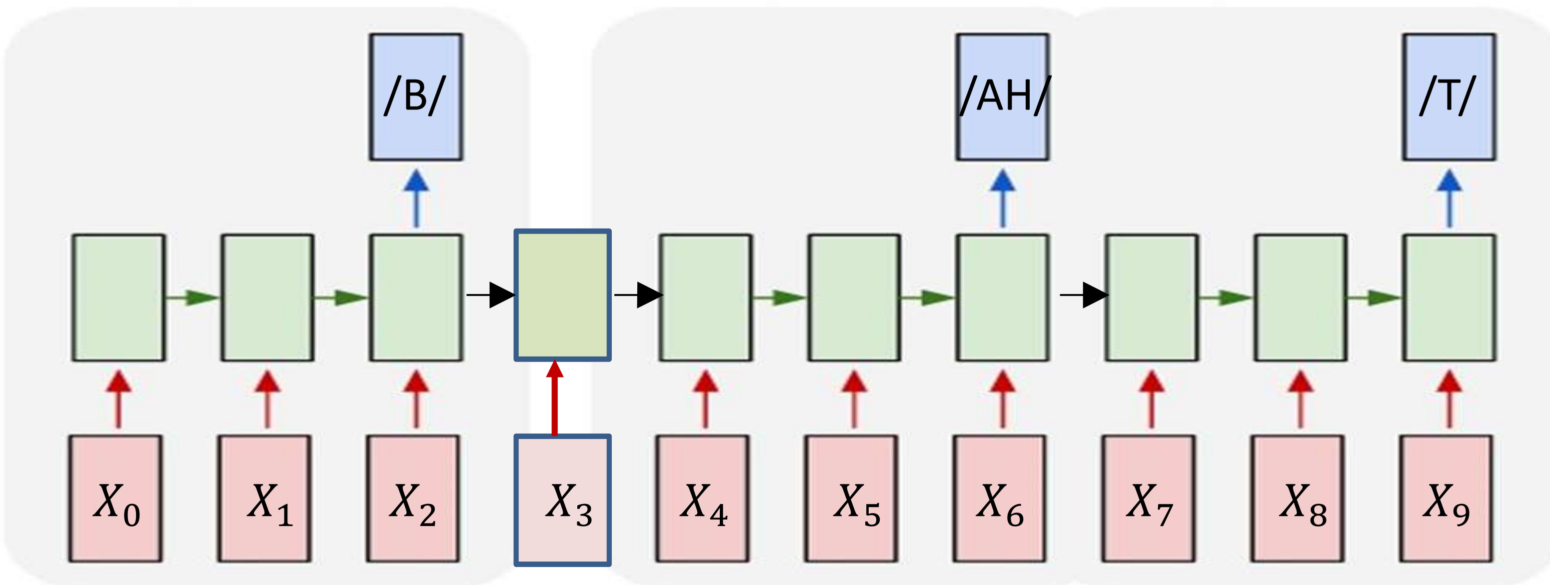


- How do we know *when* to output symbols
  - In fact, the network produces outputs at *every* time
  - *Which* of these are the *actual* outputs
- How do we *train* these models?

Partially Addressed

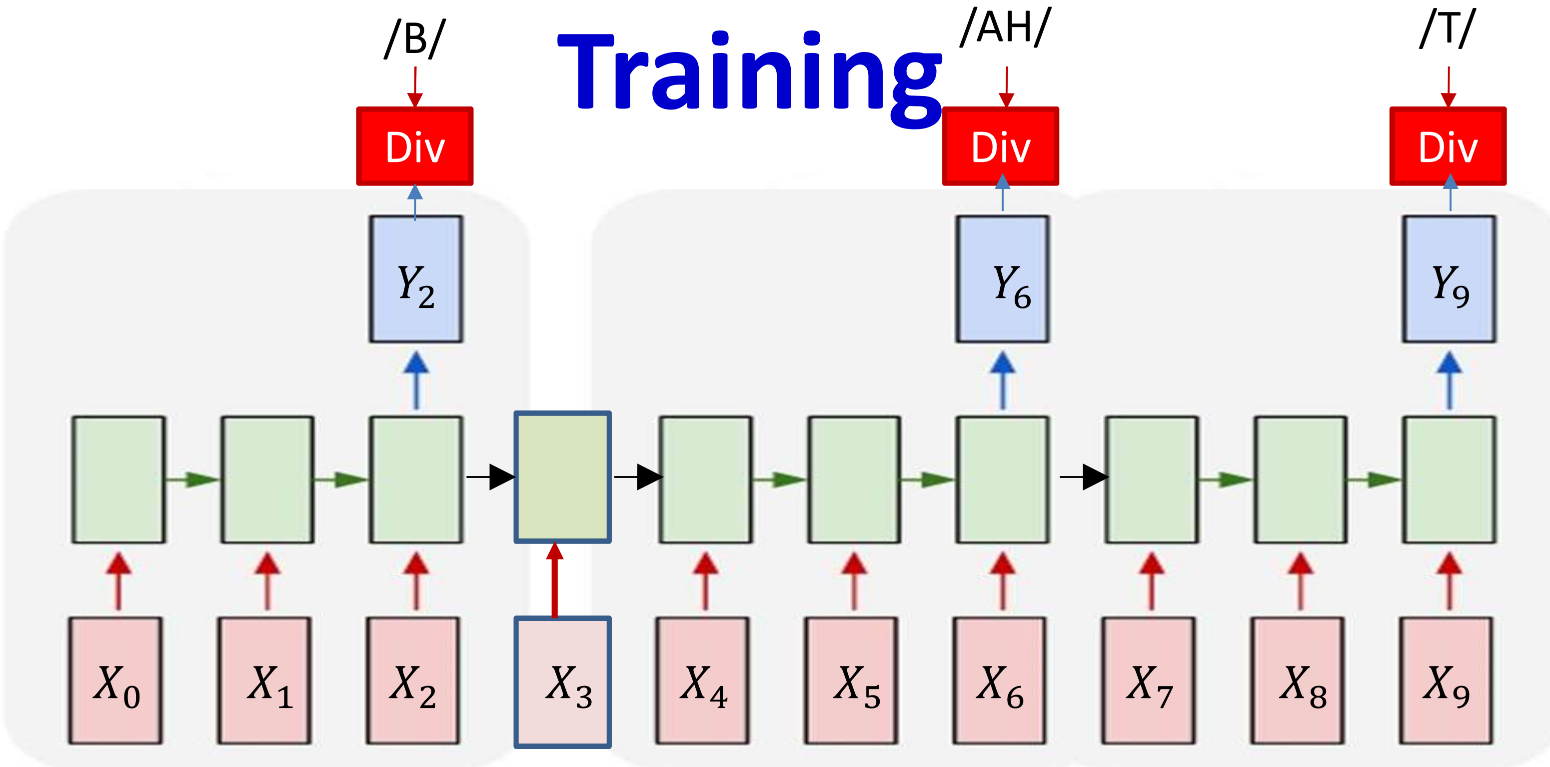
We will revisit this

Training



- Given output symbols *at the right locations*
  - The phoneme /B/ ends at X_2, /AH/ at X_6, /T/ at X_9

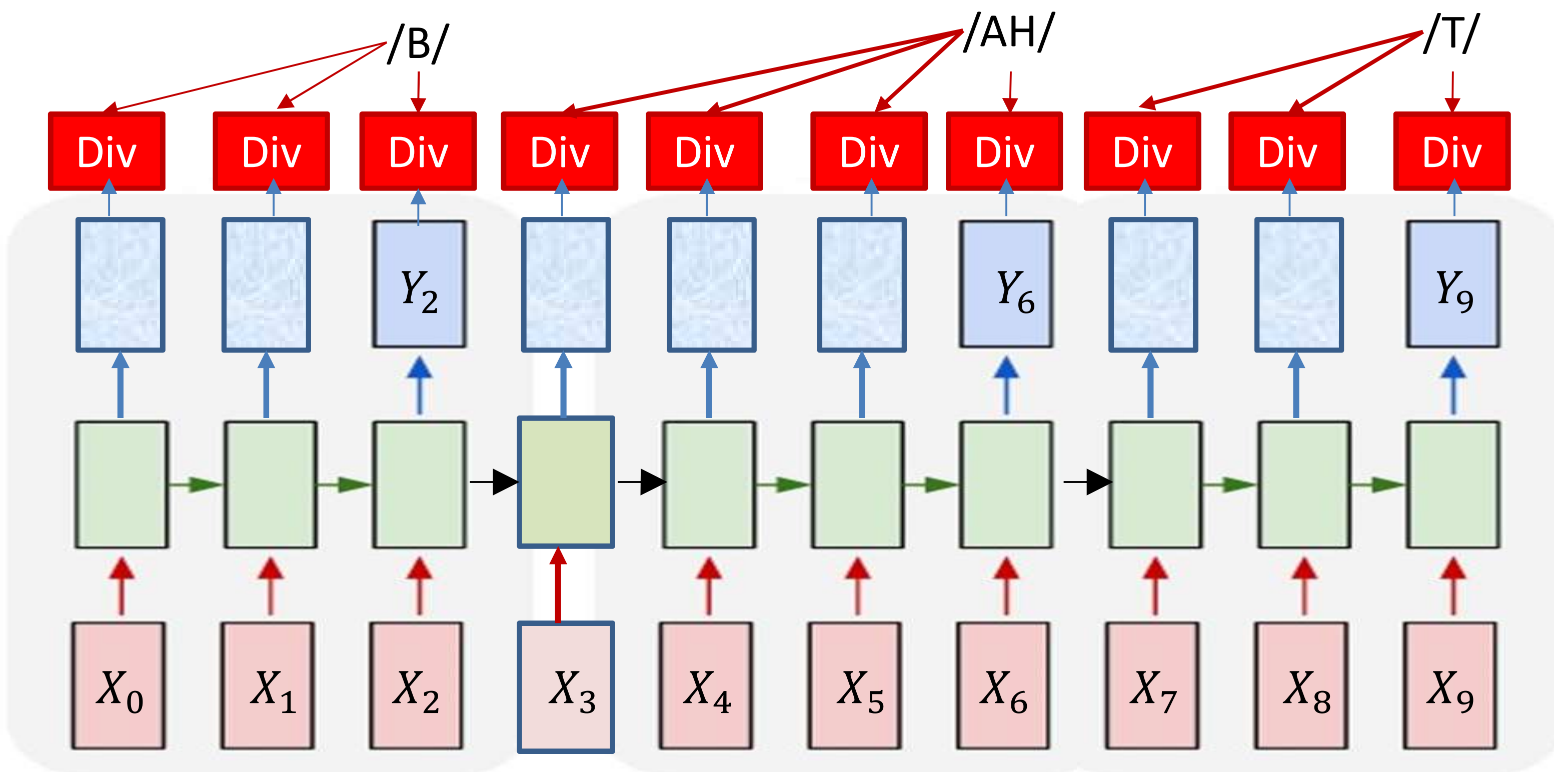
Training



- Either just define Divergence as:

$$DIV = Xent(Y_2, B) + Xent(Y_6, AH) + Xent(Y_9, T)$$

- Or..



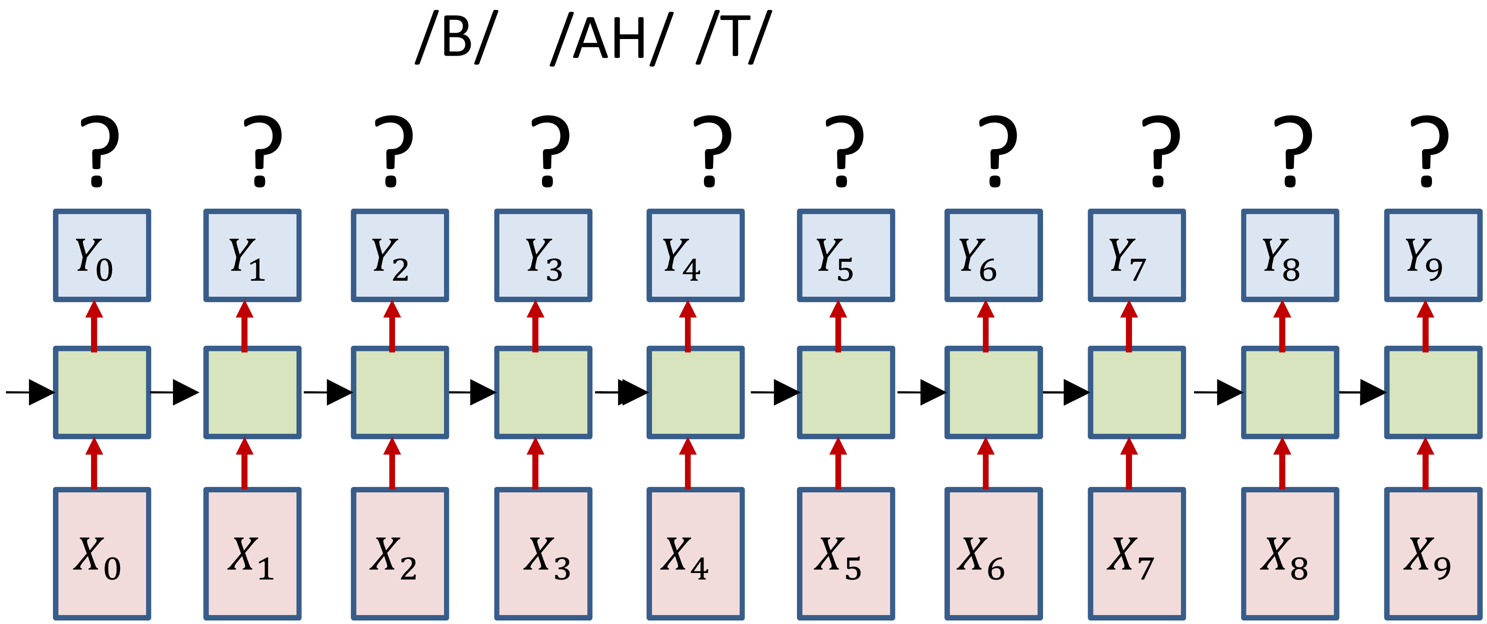
- Either just define Divergence as:

  $$DIV = Xent(Y_2, B) + Xent(Y_6, AH) + Xent(Y_9, T)$$

- Or *repeat the symbols over their duration*

$$DIV = \sum_t Xent(Y_t, symbol_t) = -\sum_t \log Y(t, symbol_t)$$

Problem: No timing information provided



- Only the sequence of output symbols is provided for the training data
  - But no indication of which one occurs where
- How do we compute the divergence?
  - And how do we compute its gradient w.r.t. Y_t

Next Class

- Training without aligned truth..
  - Connectionist Temporal Classification
  - Separating repeated symbols
- The CTC decoder..