Deep learning

4.1. DAG networks

François Fleuret

https://fleuret.org/dlc/
We can generalize an MLP
We can generalize an MLP

\[ f(x) = \sigma \left( w^{(2)} \times \sigma \left( w^{(1)} \times x + b^{(1)} \right) + b^{(2)} \right) \]

to an arbitrary “Directed Acyclic Graph” (DAG) of operators

\[ f(x) = \phi^{(1)}(x) \phi^{(3)}(\phi^{(2)}(x)) \]

François Fleuret
Deep learning / 4.1. DAG networks
Forward pass

\[
\begin{align*}
  x^{(0)} &= x \\
  w^{(1)} &\rightarrow \phi^{(1)} \rightarrow x^{(1)} \\
  w^{(2)} &\rightarrow \phi^{(2)} \rightarrow x^{(2)} \\
  \phi^{(1)} &\rightarrow x^{(1)} \rightarrow \phi^{(3)} \rightarrow f(x) = x^{(3)} \\
  \end{align*}
\]
Forward pass

\[ x^{(0)} = x \]
Forward pass

\[ x^{(0)} = x \]

\[ x^{(1)} = \phi^{(1)}(x^{(0)}; w^{(1)}) \]
Forward pass

\[ x^{(0)} = x \]
\[ x^{(1)} = \phi^{(1)}(x^{(0)}; w^{(1)}) \]
\[ x^{(2)} = \phi^{(2)}(x^{(0)}, x^{(1)}; w^{(2)}) \]
Forward pass

\[
x^{(0)} = x \\
x^{(1)} = \phi^{(1)}(x^{(0)}; w^{(1)}) \\
x^{(2)} = \phi^{(2)}(x^{(0)}, x^{(1)}; w^{(2)}) \\
f(x) = x^{(3)} = \phi^{(3)}(x^{(1)}, x^{(2)}; w^{(1)})
\]
If \((a_1, \ldots, a_Q) = \phi(b_1, \ldots, b_R)\), we use the notation

\[
\begin{bmatrix}
\frac{\partial a_1}{\partial b_1} & \cdots & \frac{\partial a_Q}{\partial b_1} \\
\cdots & \cdots & \cdots \\
\frac{\partial a_1}{\partial b_R} & \cdots & \frac{\partial a_Q}{\partial b_R}
\end{bmatrix} = J_{\phi}^T = \begin{pmatrix}
\frac{\partial a_1}{\partial b_1} & \cdots & \frac{\partial a_Q}{\partial b_1} \\
\vdots & \ddots & \vdots \\
\frac{\partial a_1}{\partial b_R} & \cdots & \frac{\partial a_Q}{\partial b_R}
\end{pmatrix}.
\]

It does not specify at which point this is computed, but it will always be for the forward-pass activations.
If \((a_1, \ldots, a_Q) = \phi(b_1, \ldots, b_R)\), we use the notation

\[
\left[ \frac{\partial a}{\partial b} \right] = J^\top_\phi = \begin{pmatrix}
\frac{\partial a_1}{\partial b_1} & \cdots & \frac{\partial a_Q}{\partial b_1} \\
\vdots & \ddots & \vdots \\
\frac{\partial a_1}{\partial b_R} & \cdots & \frac{\partial a_Q}{\partial b_R}
\end{pmatrix}.
\]

It does not specify at which point this is computed, but it will always be for the forward-pass activations.

Also, if \((a_1, \ldots, a_Q) = \phi(b_1, \ldots, b_R, c_1, \ldots, c_S)\), we use

\[
\left[ \frac{\partial a}{\partial c} \right] = J^\top_\phi|_c = \begin{pmatrix}
\frac{\partial a_1}{\partial c_1} & \cdots & \frac{\partial a_Q}{\partial c_1} \\
\vdots & \ddots & \vdots \\
\frac{\partial a_1}{\partial c_S} & \cdots & \frac{\partial a_Q}{\partial c_S}
\end{pmatrix}.
\]
Backward pass, derivatives w.r.t activations

\[ f(x) = x(3) \]

\[ \frac{\partial l}{\partial x(0)} = \frac{\partial x(1)}{\partial x(0)} \frac{\partial l}{\partial x(1)} + \frac{\partial x(2)}{\partial x(0)} \frac{\partial l}{\partial x(2)} \]

\[ \frac{\partial l}{\partial x(1)} = \phi(1)^\top \frac{\partial l}{\partial x(3)} \]

\[ \frac{\partial l}{\partial x(2)} = \phi(2)^\top \frac{\partial l}{\partial x(3)} + \phi(3)^\top \frac{\partial l}{\partial x(1)} \]

\[ \frac{\partial l}{\partial x(3)} = \phi(3)^\top \frac{\partial l}{\partial x(1)} \]

\[ \frac{\partial l}{\partial x(2)} = \phi(2)^\top \frac{\partial l}{\partial x(1)} \]

\[ \frac{\partial l}{\partial x(1)} = \phi(1)^\top \frac{\partial l}{\partial x(3)} \]
Backward pass, derivatives w.r.t activations

\[
\begin{align*}
\frac{\partial \ell}{\partial x^{(2)}} &= \left[ \frac{\partial x^{(3)}}{\partial x^{(2)}} \right] \left[ \frac{\partial \ell}{\partial x^{(3)}} \right] = \mathbf{J}_{\phi^{(3)}|x^{(2)}}^T \left[ \frac{\partial \ell}{\partial x^{(3)}} \right]
\end{align*}
\]
Backward pass, derivatives w.r.t activations

\[
\begin{align*}
\frac{\partial \ell}{\partial x(2)} & = \left[ \frac{\partial x^{(3)}}{\partial x(2)} \right] \left[ \frac{\partial \ell}{\partial x^{(3)}} \right] = J_{\phi^{(3)}|x^{(2)}}^{T} \left[ \frac{\partial \ell}{\partial x(3)} \right] \\
\frac{\partial \ell}{\partial x(1)} & = \left[ \frac{\partial x^{(2)}}{\partial x(1)} \right] \left[ \frac{\partial \ell}{\partial x(2)} \right] + \left[ \frac{\partial x^{(3)}}{\partial x(1)} \right] \left[ \frac{\partial \ell}{\partial x(3)} \right] = J_{\phi^{(2)}|x^{(1)}}^{T} \left[ \frac{\partial \ell}{\partial x(2)} \right] + J_{\phi^{(3)}|x^{(1)}}^{T} \left[ \frac{\partial \ell}{\partial x(3)} \right]
\end{align*}
\]
Backward pass, derivatives w.r.t activations

\[
\frac{\partial \ell}{\partial x(2)} = \frac{\partial x(3)}{\partial x(2)} \frac{\partial \ell}{\partial x(3)} = J_{\phi(3)}^T x(2) \frac{\partial \ell}{\partial x(3)}
\]

\[
\frac{\partial \ell}{\partial x(1)} = \frac{\partial x(2)}{\partial x(1)} \frac{\partial \ell}{\partial x(2)} + \frac{\partial x(3)}{\partial x(1)} \frac{\partial \ell}{\partial x(3)} = J_{\phi(2)}^T x(1) \frac{\partial \ell}{\partial x(2)} + J_{\phi(3)}^T x(1) \frac{\partial \ell}{\partial x(3)}
\]

\[
\frac{\partial \ell}{\partial x(0)} = \frac{\partial x(1)}{\partial x(0)} \frac{\partial \ell}{\partial x(1)} + \frac{\partial x(2)}{\partial x(0)} \frac{\partial \ell}{\partial x(2)} = J_{\phi(1)}^T x(0) \frac{\partial \ell}{\partial x(1)} + J_{\phi(2)}^T x(0) \frac{\partial \ell}{\partial x(2)}
\]
Backward pass, derivatives w.r.t parameters

\[
\begin{align*}
\partial l / \partial w^{(1)} &= \partial x^{(1)} / \partial w^{(1)} \left( \partial l / \partial x^{(1)} \right) + \partial x^{(3)} / \partial w^{(1)} \left( \partial l / \partial x^{(3)} \right) \\
\partial l / \partial w^{(2)} &= \partial x^{(2)} / \partial w^{(2)} \left( \partial l / \partial x^{(2)} \right)
\end{align*}
\]
Backward pass, derivatives w.r.t parameters

\[
\left[ \frac{\partial \ell}{\partial w^{(1)}} \right] = \left[ \frac{\partial x^{(1)}}{\partial w^{(1)}} \right] \left[ \frac{\partial \ell}{\partial x^{(1)}} \right] + \left[ \frac{\partial x^{(3)}}{\partial w^{(1)}} \right] \left[ \frac{\partial \ell}{\partial x^{(3)}} \right] = J_{\phi^{(1)}|w^{(1)}}^T \left[ \frac{\partial \ell}{\partial x^{(1)}} \right] + J_{\phi^{(3)}|w^{(1)}}^T \left[ \frac{\partial \ell}{\partial x^{(3)}} \right]
\]
Backward pass, derivatives w.r.t parameters

\[
\begin{align*}
\frac{\partial \ell}{\partial w^{(1)}} &= \left[ \frac{\partial x^{(1)}}{\partial w^{(1)}} \right] \left[ \frac{\partial \ell}{\partial x^{(1)}} \right] + \left[ \frac{\partial x^{(3)}}{\partial w^{(1)}} \right] \left[ \frac{\partial \ell}{\partial x^{(3)}} \right] = J_{\phi^{(1)} | w^{(1)}}^T \left[ \frac{\partial \ell}{\partial x^{(1)}} \right] + J_{\phi^{(3)} | w^{(1)}}^T \left[ \frac{\partial \ell}{\partial x^{(3)}} \right] \\
\frac{\partial \ell}{\partial w^{(2)}} &= \left[ \frac{\partial x^{(2)}}{\partial w^{(2)}} \right] \left[ \frac{\partial \ell}{\partial x^{(2)}} \right] = J_{\phi^{(2)} | w^{(2)}}^T \left[ \frac{\partial \ell}{\partial x^{(2)}} \right]
\end{align*}
\]

François Fleuret
Deep learning / 4.1. DAG networks
So if we have a library of “tensor operators”, and implementations of

\[
(x_1, \ldots, x_d, w) \mapsto \phi(x_1, \ldots, x_d; w)
\]

\[
\forall c, \ (x_1, \ldots, x_d, w) \mapsto J_{\phi|_{x_c}}(x_1, \ldots, x_d; w)
\]

\[
(x_1, \ldots, x_d, w) \mapsto J_{\phi|_w}(x_1, \ldots, x_d; w),
\]
So if we have a library of “tensor operators”, and implementations of

\[(x_1, \ldots, x_d, w) \mapsto \phi(x_1, \ldots, x_d; w)\]

\[\forall c, \ (x_1, \ldots, x_d, w) \mapsto J_{\phi|_{x_c}} (x_1, \ldots, x_d; w)\]

\[(x_1, \ldots, x_d, w) \mapsto J_{\phi|_w} (x_1, \ldots, x_d; w),\]

we can build an arbitrary directed acyclic graph with these operators at the nodes, compute the response of the resulting mapping, and compute its gradient with back-prop.
Writing from scratch a large neural network is complex and error-prone.
Writing from scratch a large neural network is complex and error-prone.

Multiple frameworks provide libraries of tensor operators and mechanisms to combine them into DAGs and automatically differentiate them.

<table>
<thead>
<tr>
<th>Framework</th>
<th>Language(s)</th>
<th>License</th>
<th>Main backer</th>
</tr>
</thead>
<tbody>
<tr>
<td>PyTorch</td>
<td>Python</td>
<td>BSD</td>
<td>Facebook</td>
</tr>
<tr>
<td>Caffe2</td>
<td>C++, Python</td>
<td>Apache</td>
<td>Facebook</td>
</tr>
<tr>
<td>TensorFlow</td>
<td>Python, C++</td>
<td>Apache</td>
<td>Google</td>
</tr>
<tr>
<td>JAX</td>
<td>Python</td>
<td>Apache</td>
<td>Google</td>
</tr>
<tr>
<td>MXNet</td>
<td>Python, C++, R, Scala</td>
<td>Apache</td>
<td>Amazon</td>
</tr>
<tr>
<td>CNTK</td>
<td>Python, C++</td>
<td>MIT</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Torch 7</td>
<td>Lua</td>
<td>BSD</td>
<td>Facebook</td>
</tr>
<tr>
<td>Theano</td>
<td>Python</td>
<td>BSD</td>
<td>U. of Montreal</td>
</tr>
<tr>
<td>Caffe</td>
<td>C++</td>
<td>BSD 2 clauses</td>
<td>U. of CA, Berkeley</td>
</tr>
</tbody>
</table>
Writing from scratch a large neural network is complex and error-prone.

Multiple frameworks provide libraries of tensor operators and mechanisms to combine them into DAGs and automatically differentiate them.

<table>
<thead>
<tr>
<th>Framework</th>
<th>Language(s)</th>
<th>License</th>
<th>Main backer</th>
</tr>
</thead>
<tbody>
<tr>
<td>PyTorch</td>
<td>Python</td>
<td>BSD</td>
<td>Facebook</td>
</tr>
<tr>
<td>Caffe2</td>
<td>C++, Python</td>
<td>Apache</td>
<td>Facebook</td>
</tr>
<tr>
<td>TensorFlow</td>
<td>Python, C++</td>
<td>Apache</td>
<td>Google</td>
</tr>
<tr>
<td>JAX</td>
<td>Python</td>
<td>Apache</td>
<td>Google</td>
</tr>
<tr>
<td>MXNet</td>
<td>Python, C++, R, Scala</td>
<td>Apache</td>
<td>Amazon</td>
</tr>
<tr>
<td>CNTK</td>
<td>Python, C++</td>
<td>MIT</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Torch 7</td>
<td>Lua</td>
<td>BSD</td>
<td>Facebook</td>
</tr>
<tr>
<td>Theano</td>
<td>Python</td>
<td>BSD</td>
<td>U. of Montreal</td>
</tr>
<tr>
<td>Caffe</td>
<td>C++</td>
<td>BSD 2 clauses</td>
<td>U. of CA, Berkeley</td>
</tr>
</tbody>
</table>

One approach is to define the nodes and edges of such a DAG statically (Torch, TensorFlow, Caffe, Theano, etc.)
In TensorFlow, to run a forward/backward pass on

\[ f(x) = \phi(3)(\phi(2)(\phi(1)(x_0) + w_2(x_1 + x_2)) + w_1(x_1 + x_2)) \]

\[ w_1 = \text{tf.Variable(tf.random_normal([5, 5]))} \]
\[ w_2 = \text{tf.Variable(tf.random_normal([5, 5]))} \]
\[ x = \text{tf.Variable(tf.random_normal([5, 1]))} \]
\[ x_0 = x \]
\[ x_1 = \text{tf.matmul}(w_1, x_0) \]
\[ x_2 = x_0 + \text{tf.matmul}(w_2, x_1) \]
\[ x_3 = \text{tf.matmul}(w_1, x_1 + x_2) \]
\[ q = \text{tf.norm}(x_3) \]
\[ gw_1, gw_2 = \text{tf.gradients}(q, [w_1, w_2]) \]

with tf.Session() as sess:
    sess.run(tf.global_variables_initializer())
    _gw1, _gw2 = sess.run([gw1, gw2])
In TensorFlow, to run a forward/backward pass on

\[
\begin{align*}
\phi^{(1)}(x^{(0)}; w^{(1)}) &= w^{(1)} x^{(0)} \\
\phi^{(2)}(x^{(0)}, x^{(1)}; w^{(2)}) &= x^{(0)} + w^{(2)} x^{(1)} \\
\phi^{(3)}(x^{(1)}, x^{(2)}; w^{(1)}) &= w^{(1)} (x^{(1)} + x^{(2)})
\end{align*}
\]

\[f(x) = x^{(3)}\]
In TensorFlow, to run a forward/backward pass on

\[
\begin{align*}
\phi^{(1)}(x^{(0)}; w^{(1)}) &= w^{(1)}x^{(0)} \\
\phi^{(2)}(x^{(0)}, x^{(1)}; w^{(2)}) &= x^{(0)} + w^{(2)}x^{(1)} \\
\phi^{(3)}(x^{(1)}, x^{(2)}; w^{(1)}) &= w^{(1)}(x^{(1)} + x^{(2)})
\end{align*}
\]

\[
\begin{align*}
w_1 &= \text{tf.Variable(tf.random_normal([5, 5]))} \\
w_2 &= \text{tf.Variable(tf.random_normal([5, 5]))} \\
x &= \text{tf.Variable(tf.random_normal([5, 1]))} \\
x_0 &= x \\
x_1 &= \text{tf.matmul}(w_1, x_0) \\
x_2 &= x_0 + \text{tf.matmul}(w_2, x_1) \\
x_3 &= \text{tf.matmul}(w_1, x_1 + x_2) \\
q &= \text{tf.norm}(x_3) \\
gw_1, gw_2 &= \text{tf.gradients}(q, [w_1, w_2])
\end{align*}
\]

\[
\begin{align*}
\text{with tf.Session() as sess:} \\
\quad \text{sess.run(tf.global_variables_initializer())} \\
\quad _{gw_1}, _{gw_2} = \text{sess.run([gw_1, gw_2])}
\end{align*}
\]
Weight sharing
In our generalized DAG formulation, we have in particular implicitly allowed the same parameters to modulate different parts of the processing.
In our generalized DAG formulation, we have in particular implicitly allowed the same parameters to modulate different parts of the processing.

For instance $w^{(1)}$ in our example parametrizes both $\phi^{(1)}$ and $\phi^{(3)}$.

This is called weight sharing.
Weight sharing allows in particular to build **Siamese networks** where a full sub-network is replicated several times.
Weight sharing allows in particular to build **Siamese networks** where a full sub-network is replicated several times.
Weight sharing allows in particular to build **Siamese networks** where a full sub-network is replicated several times.
Weight sharing allows in particular to build **Siamese networks** where a full sub-network is replicated several times.
Weight sharing allows in particular to build **Siamese networks** where a full sub-network is replicated several times.
The end